Brandenburg University of Applied Sciences

IT Security
Computerscience
Prof. Dr. Oleg Lobachev
MSc. Florian Eich

Applied Graph Kernels for Schema-Aware In-Context Learning in NL2SQL

Bachelor Thesis

Winter Semester 2025

January 26, 2026

Mara Schulke – Matr-Nr. 20215853

Abstract

This thesis explores the integration of large language models (LLMs) into PostgreSQL database systems in order to make the database accessible via natural language instead of the postgres SQL dialect. The research focuses on implementation strategies, performance optimization, and practical applications of this concept.

Contents

List of Figures

List of Abbreviations

GPT	Generative Pretrained Transformer
SQL	Structured Query Language
API	Application Programming Interface
LLM	Large Language Model
DBMS	Database Management System
NL2SQL	Natural Language to SQL

1 Introduction

1.1 Problem Statement and Motivation

Database systems represent a backbone of modern computer science, allowing for rapid advancements whilst shielding us from the problem categories that come along with managing and querying large amounts of, usually structured, data efficiently. However, most Database Management Systems (DBMS) have traditionally required specialized knowledge, usually of the Structured Query Language (SQL), in order to become useable. Whilst this barrier may be percieved differently across diverse usergroups it represents a fundamental misalignment between end-user goals (e.g. analysts, researchers, domain experts etc.) and the underlying DBMS, thus often requiring software engineering efforts in order to reduce this friction.

This barrier is the reason entire classes of software projects exists (for example, admin / support panels), data analytics tools etc. which therefore introduce significant churn and delay between the implementation of a database system and reaching the desired end user impact. Often these projects span multiple years, require costly staffing and yield little to no novel technical value.

Emerging technologies such as Large Language Models (LLMs) have proven themselves as a sensible tool for bridging fuzzy user provided input into discrete, machine readable formats. Prominent models in this field have demostrated outstanding capabilities that enable computer scientists to tackle new problem classes, that used to be challenging / yielded unsatisfying results with logical programming approaches.

This thesis is exploring ways to overcome the above outlined barrier using natural language queries, so that domain experts, business owners, support staff etc. are able to seamlessly interact with their data, essentially eliminating the requirement of learning SQL (and its pitfalls). By translating natural language to SQL using Large Language Models this translation becomes very robust (e.g. against different kinds of phrasing) and enables novel applications in how businesses, researchers and professionals interact with their data — it represents a fundamental shift (ie. moving away from SQL) towards a more inclusive and data driven world.

1.2 Objectives of the Thesis

This thesis aims to address the aforementioned challanges when it comes to database accessibility. The following objectives are the core research area of this thesis:

1. Develop a portable and offline-only NL2SQL system that can translate natural language queries into semantically accurate SQL queries using Large Language Models.
2. Evaluate the effectiveness and feasibility of different Models aswell as machine learning techniques in order to improve the performance of the system.
3. Benchmark the performance of the implementation against common NL2SQL benchmarks.
4. Idenitfy potential use cases for real world scenarios that could deliver a noticable improvements to users.
5. Analyze the short commings and limitations of the developed system and propose potential solutions to overcome them.

1.3 Research Questions

While recent NL2SQL research has achieved outstanding results using proprietary large language models like GPT-4 and Claude, these systems introduce significant data privacy and security concerns for real world adoption: All database schemas and queries must be transimitted to an external API that is used for the actual inference. This incurrs risks, ongoing and uncontrolled costs and a dependency on a third-party service being available. This thesis investigates the applicability and competitiveness of open-source NL2SQL models deployed locally on consumer hardware. Therefore the following research questions are addressed:

RQ1 — To what extent can open-source LLMs achieve competitive NL2SQL performance through pipeline composition and optimization?

State-of-the-art systems often combine multiple NLP and ML techniques like in-context learning, self-correction and candidate selection. Often they rely on closed-source models which provide impressive baseline performance with state-of-the-art reasoning capabilities. This research question investigates whether the emerging gap between proprietary systems and open-source models like -7B can be reduced or closed by combining the models with other NLP or ML techniques. Specifically:

- Can pipeline composition improve the performance of fine-tuned models significantly?
- How does the performance of open-source NL2SQL systems compare to closed-source model baselines on standard benchmarks?
- What accuracy ceiling can be achieved within consumer hardware constraints?

RQ2 — How do NL2SQL pipeline components interact, and which configurations optimize the accuracy-latency tradeoff?

In order to resolve the dependency on proprietary systems for NL2SQL, a cost effective local deployment must be achieved. Thus an offline NL2SQL system must utilize compute efficiently as simply combining all available techniques may not be optimal for performance characteristics. If components have dependencies, redundancies, or unfavorable cost-benefit ratios the overall accuracy-latency ratio would be reduced. This research question investigates the interactions of pipeline components and which configurations are favorable:

- Do independent components of NL2SQL systems provide independent benefits or do they interact synergistically?
- Which components contribute most to accuracy improvement, and which add latency without commensurate benefit?
- How should pipelines be configured for different deployment scenarios (exploratory analysis vs production queries, simple vs complex databases)?
- What is the optimal point for the accuracy-latency tradeoff for different use cases?

1.4 Structure of the Thesis

This thesis is following a research and development methodology in order to implement a portable NL2SQL system that can be used in sovereign environments and be embedded into databases.

1. **Literature Review** — An analysis of the existing research in the fields of natural language interfaces (NLI) for databases, GPU integration for acceleration of database operations, and LLM/AI Model integration with database systems. This phase establishes the theoretical foundation for this research and identifies current state-of-the-art approaches, their benefits and shortcomings.

2. **Theoretical Foundations** — Introducing the theoretical concepts and frameworks required for understanding the problem statement and the proposed solution. The fundamentals introduced in this section are referenced in the system design and implementation phases respectively.

3. **System Design** — Design of a system architecture that can translate natural language to SQL using LLMs. The primary goal of the system design phase is to arrive at an architecture that yields low latency natural language processing, schema-aware SQL query generation and ambiguity resolution whilst maintaining a high semantic accuracy.

4. **Implementation** — The implementation of a PostgreSQL extension according to the above system design that relies on `rust` and `pgrx`. This extension will provide a GPU accelerated framework for executing LLMs, implement a natural language to query generation pipeline that relies on the SQL schema and create database functions and operators for both query generation and execution.

5. **Evaluation** — Benchmarking the implementation against standard evaluation datasets framework and benchmark that introspects the implementations performance in multiple dimensions. The most relevant evaluation metrics for this thesis are execution accuracy ($\mathbb{EA}$), exact match ($\mathbb{EM}$), error rate ($\mathbb{ER}$) and candidate latency ($\mathbb{CL}$).

6. **Discussion** — Analysis and interpretation of the evaluation phase results against the research goals of this thesis. Evaluating the performance and accuracy results recorded during the benchmarks against the question whether real world deployments of NILs are feasible. Furthermore the impact of approaches used are shown and it is determined whether a statistically significant improvement can be achieved.

7. **Summary and Outlook** — Summarizes the contributions, addresses limitations of this thesis and the implementation, and proposes directions for future research alongside possible applications. Primary future research topics include advanced GPU optimization techniques (e.g. further quantization), accuracy and performance impact of model fine tuning, techniques, scalability of such a system in enterprise scenarios and the evaluation of security and privacy considerations (e.g. managing access control).

2 Literature Review

In this section a comprehensive literature review is performed to asses the research landscape on NL2SQL (sometimes also referred to as Text-to-SQL or T2SQL) and NLIDBs. From the time their development accelerated in the late 1990s and early 2000s (Androutsopoulos, Ritchie, & Thanisch, 1995; Popescu, Etzioni, & Kautz, 2003; Tang & Mooney, 2001; Zelle & Mooney, 1996) until now, observing multiple larger paradigm shifts happening over time (Deng et al., 2020; F. Li & Jagadish, 2014; Yaghmazadeh, Wang, Dillig, & Dillig, 2017; Yu et al., 2020; Zhong, Xiong, & Socher, 2017). In particular this research focuses on the recent advancements when it comes to language models and how they can be harnessed for effective NL2SQL systems (D. Gao et al., 2023; Lei et al., 2025; J. Li, Hui, Qu, et al., 2023; Rahaman, Zheng, Milani, Chiang, & Pottinger, 2024; Rajkumar, Li, & Bahdanau, 2022; B. Zhang et al., 2024).

This literature review is covering the foundational concepts, challenges, key advancements and research gaps associated with using natural language instead of SQL. It lays the foundation for this thesis and helps to set the research questions introduced in the previous chapter in context.

2.1 Foundations of Natural Language Interfaces to Databases

Earlier papers in the research landscape on Natural Language Database Interfaces (NLIDBs) date over half a century back, into the early 1970s. Two decades after the first major research systems where developed in this domain, Androutsopoulos, Ritchie, and Thanisch have published an introduction and an overview over NLIDBs where an overview of state-of-the-art approaches were provided. (Androutsopoulos et al., 1995) Their work outlined multiple key issues and challenges associated with NLIDBs, and compared them against existing / competing solutions like formal query languages, form-based interfaces and graphical interfaces. These challenges (like unobvious limits, linguistic ambiguities, semantic inaccuracy, tedious configuration etc.) have shaped this field of research and are still considered relevant metrics today.

Early NLIDBs primarily relied on traditional natural language processing (NLP) techniques in order to achieve natural language understanding capabilities. With Chill an inductive logic programming (ILP) approach was first introduced for NL2SQL systems, marking one of the key events when it comes to machine learning usage. (Zelle & Mooney, 1996) In 2001 Tang and Mooney have extended the approach of ILP parsing for natural language database queries with multi clause construction, yielding promising results in the field of NLIDBs. (Tang & Mooney, 2001)

Building on the systematic overview of Androutsopoulos, Ritchie, and Thanisch and the first machine learning approaches from Zelle and Mooney aswell as Tang and Mooney, Popescu et al. have proposed a novel approach for implementing NLIDBs and outperformed at the time state-of-the-art solutions from Zelle and Mooney (1996) Tang and Mooney (2001) — achieving 80% semantic accuracy. (Popescu et al., 2003) The novelty of the Percise system lies in its natural language processing approach, specifically its lexical mapping strategy, allowing Percise to identify questions it can, and can't answer (introducing the concept of *semantically tractable questions*) which therefore results in a better and interactive end user experience. Their experiements also showed that this approach is *transferrable* and *unbiased* — it is possible to feed in new, unknown questions into the system and maintain performance characteristics, whereas it was shown that Zelle and Mooney (1996) were suffering from a distribution drift of the questions asked. (Popescu et al., 2003)

The theoretical foundations and research questions highlighted by the aforementioned works, shaped the research field and highlighted the following, ongoing research:

1. The trade-off characteristics derived from choosing a machine learning vs. traditional NLP approach (e.g. Chill versus Percise). E.g. coverage versus correctness. (Popescu et al., 2003; Zelle & Mooney, 1996)

2. The linguistic challenges associated with bringing NLIDBs into use (e.g. semantic inaccuracy, linguistic ambiguity, unclear language coverage etc.) (Androutsopoulos et al., 1995)

3. The value of systems and approaches which double down on reliability and semantic accuracy rather than giving promising but incorrect answers. (Androutsopoulos et al., 1995; Popescu et al., 2003)

Fundamentally this highlights the tension and mismatch between the characteristics of natural language, which is able to be ambiguous, *semantically untractable* or able to be incomplete in meaning and formal languages like SQL which always have on deterministic and *semantically tractable* meaning they convey in each statement. As Schneiderman and Norman have pointed out according to Popescu, Etzioni, and Kautz, users are "unwilling to trade reliable and predictable user interfaces for intelligent but unreliable ones" which induces performance

expectations on NLIDB implementations to be highly certain about the questions it can, and can't answer, whilst maintaing as high as possible natural language coverage. (Popescu et al., 2003)

2.2 Traditional NL2SQL Approaches

Prior to the wide-spread dominance of machine learning approaches for natural language processing a variety of traditional, rather discrtete approaches have been explored in the field of NL2SQL / NLIDBs. These logical programming approaches have laid the foundations for transitioning towards the application of machine learning techniques for NL2SQL.

2.2.1 Rule-based and Grammar-based Systems

Foundational research of NL2SQL system mostly focused around applying rule engines that were tedious to set up and expensive to maintain / transfer across database systems. These rule engines mostly relied on the systematic identification of linguistic patterns / were trying to template SQL from information that was derived from processing the natural language query. (Codd, 1974; Hendrix, Sacerdoti, Sagalowicz, & Slocum, 1978; Woods, Kaplan, & Nash-Webber, 1972) These approaches mostly tried to formalize natural language queries into formal grammers which could then be deterministically mapped into a valid SQL query. (Woods et al., 1972) These approaches have strong downsides when it comes to the variety of natural language constructs they can process, aswell as runtime adoption of new / unknown databases, query constructs etc. A potential upside of this class of NL2SQL systems is that they can confidently and reproducably identify questions they can, and can't answer — thus leading to very reliable and predictable user interfaces.

2.2.2 Semantic Parsing using String-Kernels

A significiant milestone in parsing techniques of natural language queries was reached by Kate and Mooney in 2006. The introduction of string kernels for semantic parsing represented a novel achievement, when it comes to fusing logical programming approaches using a formal grammer like LSNLIS developed by Woods et al. (1972) and learning / training approaches to understand unseen language patterns / unknown natural language query structures. This allowed for more flexible pattern recognition when compared to traditional rule-based systems.

The core innovative characteristic of this approach lies in its capability to understand similarities between natural language expressions based on subsequence patterns rather than relying on exact matches. This made Krisp, the research NLIDB system developed by Kate and Mooney (2006) much more robust to language variations in phrasing and noise (e.g. spelling mistakes) in the input. As the Kate and Mooney demonstrated through experiments on real-world datasets, this approach compared favorably to existing systems of the time like Chill, especially in handling noisy inputs — a frequent challenge rigid rule-based systems faced in real world scenarios (Kate & Mooney, 2006; Zelle & Mooney, 1996).

2.2.3 Graph Matching Methods

Reddy, Lapata, and Steedman (2014) brought together several research threads and reapplied emerging graph matching research models to natural language processing, specifically to natural language queries. Graph matching was applied once the natural language query was parsed using a Combinatory Categorial Grammar (CCG) approach into a semantic graph which denotes the relationship between semantic entities in it. This graph could then be matched against the actual graph derived from the database, since they share topological traits that can be used for matching (Reddy et al., 2014). This approach allowed to apply querying systems without having any question-answer pairs or manual annotations for training the system, which implies easier scalability / transferability across domains, since the system does not require any additional tweaks.

Even though this approach was novel and showed improved the performance over existing state-of-the-art approaches, it was showing that graph matching quickly reaches its limitations. This approach relied heavily on the CCG parser's accuracy, with parsing errors accounting for 10-25% (depending on the dataset) of system failures (Reddy et al., 2014). Furthermore it struggled with both ambiguous language constructs and potential mismatches between natural language representation of relationships and database layouts — more complicated database designs, which may not match the users intuitive understanding resulted in a different topology and hence could not be matched (Reddy et al., 2014, p. 387).

2.2.4 Interactive Systems

In 2014 F. Li and Jagadish identified that perfect translation of natural language into SQL was challenging due to natural language not being made for query expressions as it heavily relies on contextual information and clarifying questions in order to disambiguate conversations (F. Li & Jagadish, 2014). These learnings relate to early prior art from Montgomery and Codd which also made this observation — "natural language is not a natural query language." (Montgomery, 1972). The solution introduced by NaLIR further emphasized how important an interactive, conversational usage model is, when offering a natural language interface (F. Li & Jagadish, 2014).

NaLIR could accept logically complex English language sentences as input and translate them into SQL queries with various complexities, including aggregation, nesting, and different types of joins etc. The key innovative characteristic of NaLIR lies in its interactive communication mechanism (much like Rendezvous) that could detect potential misinterpretations and engage users to resolve ambiguities present in their natural language query without forcing them to entirely rephrase their query F. Li and Jagadish (2014). This approach, while showing awareness for its limitations (with regards to entirely automating / deriving SQL generation from potentially ambiguous or faulty user input) showed that it was possible to overcome these limtations through choosing the right interaction model — "In our system, we generate multiple possible interpretations for a natural language query and translate all of them in natural language for the user to choose from" —, rather than optimizing the generation part of the system F. Li and Jagadish (2014).

2.2.5 Query Synthesis

Yaghmazadeh et al. (2017) introduced SQLizer, which synthesizes SQL queries from natural language (Yaghmazadeh et al., 2017). This paper presents a novel approach when it comes to NL2SQL as it is merging prevalent semantic parsing techniques (outlined above) with an program synthesis (or query synthesis) approach. SQLizer makes use of a three stage processing model for natural language models: first generating a sketch of the query using semantic parsing, then using type-directed synthesis to complete the sketch and finally using automated repair, if required.

Yaghmazadeh, Wang, Dillig, and Dillig show that alternating between repairing and synthesis yields results that beat state-of-the-art NL2SQL approaches like NaLIR. SQLizer is fully automated and database-agnostic, requiring no knowledge of the underlying schema. The authors evaluated SQLizer on 455 queries across three databases, where it ranked the correct query in the top 5 results for roughly 90% of the queries. This represents a significant improvement over NaLIR (F. Li & Jagadish, 2014), the previous state-of-the-art system (Yaghmazadeh et al., 2017).

Potential short commings of this approch include queries which yield empty results, dealing with language variations as SQLizer is still using semantic parsing, and domain-specific terminology, all while still requiring users to select from multiple query options which reduces the overall usability of the system (Yaghmazadeh et al., 2017, p.22-23).

2.2.6 Limitations of Traditional Approaches to NL2SQL

Despite being innovative and achieving state-of-the-art results, many of the above outlined approaches face severe challenges when moving outside of an research environment. Many of these systems performed comparatively good on research benchmarks that were often composed of controlled question types and limited data variety. Ultimatively no standard benchmark existed for NL2SQL in this era, hence comparing different NL2SQL systems against each other is a problem on its own. Despite not having a standard benchmark that all approaches could be unifiably evaluated against, several fundamental challenges emerged / remained with these approaches:

1. **Limited linguistic coverage** — Prevalent rule-based and semantic-parsing based systems were only able to process the a small subset of the natural language they were programmed for. This severely limited their ability to handle different phrasings of the same end-user goal (Hendrix et al., 1978; Kate & Mooney, 2006; Montgomery, 1972; Woods et al., 1972).

2. **Transferability** — Traditional approaches typically required extensive manual configuration or at least a training phase / adaption for each database they were deployed for, hindering cross domain usage through being expensive and time-consuming to adapt (Androutsopoulos et al., 1995; Woods et al., 1972).

3. **Brittleness** — Many of the systems introduced in this subchapter did not handle synonyms, paraphrasing, or spelling errors well. Manual adaption / handling was needed in order to becomes resilient against each class of problems (Kate & Mooney, 2006; Yaghmazadeh et al., 2017).

4. **Poor scalability** — With potentially more complex underlying databases, traditional solutions often showed to perform worse. Reddy, Lapata, and Steedman found, that with increasing schema complexity more compute was required to resolve the natural language query to a suitable query candidate making them less transferable and scalable than initially anticipated (Reddy et al., 2014) — "Evaluating on all domains in Freebase would generate a very large number of queries for which denotations would have to be computed ... Our system loads Freebase using Virtuoso and queries it with SPARQL. Virtuoso is slow in dealing with millions of queries indexed on the entire Freebase, and is the only reason we did not work with the complete Freebase." which indicates underlying system design issues with runtime complexity.

These flaws of traditional NL2SQL approaches made it apparent, that a different class of approaches is needed, which increase transferability and reduce the brittleness since users are "unwilling to trade reliable and predictable user interfaces for intelligent but unreliable ones" according to Popescu et al. (2003). Whilst many approaches outlined tractable ways to increase user satisfaction and accuracy (like Codd did in 1974 with a conversational approach), NLIDBs were and are not considered to be a solved problem.

2.3 Neural NL2SQL Approaches

The previously outlined limitations of traditional approaches to solving NL2SQL / implementing NLIDBs pushed the research branch around neural network application forward to step in and propose new solutions which address the brittleness, transferability and scalability concerns addressed with logical programming approaches. Neural approaches showed to yield significant improvements in terms of transferability and overall accuracy which led to a paradigm shift in this research field.

2.3.1 Early Neural Approaches

In 2017 Zhong, Xiong, and Socher released Seq2SQL which represents a significant breakthrough and leap in NLIDB research. Seq2SQL was an early research system that in the field of neural network application and as one of the first papers to frame the implementation of NLIDBs / NL2SQL Systems as a reinforement learning problem. The system utilized iterative query execution in the reward function to improve its accuracy (Zhong et al., 2017). In the same paper Zhong, Xiong, and Socher introduced WikiSQL, a training dataset, which enables large scale (in 2017) model training.

SQLNet (Xu, Liu, & Song, 2017) addressed primarily the order-sensitivitiy trait of Seq2SQL (Zhong et al., 2017) that was prevalent due to being a derivative approach from sequence-to-sequence approaches. SQLNet diverges from sequence-to-sequence and joins multiple research threads, employing a sketch-based query generation. SQLNet breaks down complex queries into smaller (hence more manageable) sub-queries which can then be individually sketched and refined, yielding a system that outperformed state-of-the-art by 9% to 13% (Xu et al., 2017).

Yu, Li, Zhang, Zhang, and Radev have introduced TypeSQL, a variation of the SQLNet-approach, in 2018. TypeSQL's primary difference to SQLNet is the encoding of type information for SQL generation. The approach scanned for entity references and values in natural language and was able to improve performance by 5.5% over SOTA-Models like SQLNet whilst requiring significantly less training time, indicating that type information was a useful information for deriving accurate SQL queries from user input (Yu, Li, et al., 2018).

2.3.2 Intermediate Neural Developments

Later in 2018 Yu, Yasunaga, et al. released SyntaxSQLNet, a followup research to TypeSQL, which represented a slight change in approach and research focus. In direct comparison SyntaxSQLNet focused primarily around complex query generation using a syntax tree decoder, allowing for longer and more cohesive query generation (Yu, Yasunaga, et al., 2018). This advancement over TypeSQL allowed more complex queries to be reliably generated, enabling multiple clauses aswell as nested queries. SyntaxSQLNet was one of the earlier research efforts which utilized Spider instead of WikiSQL (introduced by Zhong et al. (2017)), a large-scale NL2SQL dataset, incorperating 10.181 hand annotated natural language question and alongside 5.693 unique SQL examples that spread across 138 different domains (Yu, Zhang, et al., 2018). This research led the transition of

comparatively simple, research-grade, neural systems for NLIDBs towards systems which are feasible in the real world.

Building on the above approaches, Guo et al. have introduced IRNet, a neural network approach using intermediate representation as a bridge between natural language and SQL in which semantic queries could be expressed. The intermediate format SemQL (or semantic query language) was utilized to transform and synthesize queries on the actual database schema more accurately than Seq2SQL. IRNet followed a three phase approach: schema linking between the natural language query and database layout, synthesis of SemQL as intermediate representation and deterministic conversion of SemQL to SQL. This approach allowed IRNet to outperform state-of-the-art approaches on the SPIDER benchmark by 19.5%, placing IRNet at an overall accuracy of 46.7% (Guo et al., 2019).

Following IRNet, graph neural networks (GNN) have been explored as alternative architecture by Bogin, Berant, and Gardner (2019), representing the database schema as a graph and using message passing to model relationships between tables, columns and natural language input. This approach demonstrated the capability to improve reasoning and query generation capability. Bogin et al. showed that when evaluating against the SPIDER benchmark GNN outperforms both SyntaxSQLNet (and therefore SQLNET and TYPESQL). Although presenting a signficiant advancement over previous state-of-the-art approaches, GNN falls behind in performance against IRNet by 6% (Bogin et al., 2019; Guo et al., 2019).

2.3.3 Relation-Aware Transformer Approaches

The release of RAT-SQL (Relation-Aware Transformer for SQL) Wang, Shin, Liu, Polozov, and Richardson (2020) represents the most significant leap in research of neural NL2SQL approaches. RAT-SQL diverged from earlier research through emphasizing the relationship between natural language and the database schema elements using relation-aware self-attention representing a novel approach for solving *schema linking* (Wang et al., 2020).

RAT-SQL's primary innovations was the ability to infer, understand and utilize the relationship between individual tokens in the natural language query and link it to the database schema. Thus allowing for reasoning capabilities on the actual database schema while generating the query.

This architecture yielded a 57.2% in exact match accuracy when being evaluated on the SPIDER benchmark, substantially outperforming comparative approaches like GNN, IRNet and IRNet V2 by 10.5%, 9.8% and 8.7% respectively. Although overall accuracy improved across all approaches when being paired with BERT (Bidirectional Encoder Representations from Transformers, a popular pre-trained lanugage model from Google) the δ between the indivudual approaches remained relatively steady, leaving RAT-SQL outperforming state-of-the-art approaches by 5% to 12.2% further demonstrating the capability advancement yielded by this system (Wang et al., 2020).

2.3.4 Comparative Analysis of Neural Approaches

The evolution from early neural approaches to RAT-SQL emphasized the rapid advancements that happened in the research field of neural NL2SQL approaches in different dimensions:

1. **Model Complexity** — Given the research progression from early sequence to sequence translation approaches (SEQ2SQL) towards sketch based and type augumented and graph based approaches (TYPESQL, SQLNET, GNN) and syntax tree decoding emphasized by SYNTAXSQLNET, neural approaches continously advanced in the complexity of approaches that is required to beat state-of-the-art approaches in contemprorary benchmarks like SPIDER. RAT- SQL presents one of the late and most complex advancements in the field of neural NL2SQL approaches with its adapted self attention mechanism (Bogin et al., 2019; Wang et al., 2020; Yu, Li, et al., 2018; Yu, Yasunaga, et al., 2018; Zhong et al., 2017).

2. **Transferability** — Each of the approaches introduced above represents a succession in terms of their transferability. The field of neural NL2SQL approaches significantly improved the ability for NLIDBs to generalize over the underlying database schemas. RAT-SQL showed the strongest cross-domain accuracy (that is benchmarked by the SPIDER benchmark). With standard benchmarks emerging it became easier to verify and quantify which approach had the highest transferability as SPIDER specifically had independent development and test datasets, preventing approaches from over-optimizing on training data (Wang et al., 2020).

3. **Robustness** — As research systems advanced in complexity and shifted from raw input to output translation (Seq2SQL) their robustness steadily increased. Through more approaches like SyntaxSQLNet which utilized structured decoding, IRNet which relied on an intermediate representation and RAT-SQL the challenges around *schema linking* outlined by Wang et al. (2020) have increasilingly led to more robust systems that can handle rephrasings, spelling mistakes and variations in natural language usage far beyond what traditional NL2SQL approaches could accomplish Guo et al. (2019); Wang et al. (2020); Yu, Yasunaga, et al. (2018).

4. **Query Complexity** — The performance on complex queries involving multiple tables, relying on complex aggregations, nested structures and joins dramatically improved over the course of the research that happened in this field. Whilst IRNet reresents one of the first singificant advancements when it comes to the ability of neural approaches to handle complex queries, RAT-SQL still showed to outperform the intermediate representation approach introduced by IRNet by up to 10.5% (Guo et al., 2019; Wang et al., 2020).

5. **Schema Understanding** — Whilst early approaches like Seq2SQL primarily applied reinforcement learning for end to end query generation (Zhong et al., 2017), later approaches like TypeSQL, GNN and specifically RAT-SQL showed novel and state-of-the-art *schema understanding / schema linking* capabilities, yielding the ability to accurately reason about user intent and traverse the database schema while generating queries (Bogin et al., 2019; Wang et al., 2020; Yu, Li, et al., 2018).

2.3.5 Limitations of Neural Approaches

Despite the dramatic *accuracy, transferability* and *robustness* improvements that could be observed with late neural approaches (Guo et al., 2019; Wang et al., 2020), neural approaches still suffered from serious shortcommings / unsolved challenges:

1. **Training Data** — Utilizing neural networks these approaches required susbtantially more training data (ie. natural language paired with output SQL queries) than traditional systems which required serious efforts of data collection (Yu, Zhang, et al., 2018).

2. **Correctness** — The inherent mismatch between neural networks and formal languages yielded cases where models produced invalid SQL code. Approaches like SyntaxSQLNet improved the tried to solve this circumstance by utilizing syntax trees during decoding but nonetheless syntactic correctness remained a challenge across future iterations of neural systems. (Yu, Yasunaga, et al., 2018)

3. **Domain Language** — Despite increased *transferability* characteristics neural approaches still suffered from a limited vocabulary and inter-domain understanding of terminology and relation between concepts which made highly domain specific natural language queries challenging.

4. **Observability** — The black-box nature of neural networks made approaches relying on them, particularly the ones with complex architectures, hard do understand / explain in case when neural systems yielded undesirable output.

The introduction and advancement of early neural NL2SQL approaches led to significant advancements in the research and feasibility of NLIDBs. The research shift started in this era established the foundations for further and more advanced machine learning approaches (specifically language model oriented approaches)being researched. Neural approaches showed to significant improvements in performance when being paired with pre-trained language models (Wang et al., 2020) which led to further research on their applicability.

2.4 Pre-trained Language Models

The advantages of combining specialized neural networks with general-purpose pre-trained language models led to a pivotal point in the NL2SQL research field towards focusing increasingly on the application of pre-trained language models for NLIDBs. Models like BERT or T5 offer noticable performance improvements (especially when it comes to language understanding) over specialized NL2SQL networks due to training happening on unrestrained amounts of natural language data, instead of pure NL2SQL datasets which are often fairly limited in size and therfore natural language use — Spider2.0 which is a contemprary NL2SQL benchmark consists of just 632 real-world questions (Lei et al., 2025). Thus PLM-based (or at least augumented) NL2SQL systems

can observe dramatic performance improvements through the language models's ability to understand patterns and identify semantic relationships of natural language query elements.

2.4.1 Early Pre-trained Language Model Adaptations

The above outlined benefits have led to concrete research efforts focusing on the question whether the sole application of pre-trained language models could outperform neural state-of-the-art approaches — which often implicitly require a far more sophisticate architecture when it comes to natural language analysis.

In the time of emerging PLM applicationm GRAPPA was introduced by Yu et al. in 2020 — a novel grammar-augmented pre-training approach built on $\text{RoBERTa}_{\text{LARGE}}$ (a derivative model from BERT). It generates synthetic training data (ie. natural language and sql pairs) using a synchronous context-free grammar (SCFG) which analyses and identifies patterns in natural language queries that can be used as templates for sythesizing training data. The specialized pre-training helps GRAPPA to establish a robust connection between natural-language and database schema elements, showing significant improvements on existing approaches on multiple contemporary benchmarks like SPIDER and WIKISQL (Yu et al., 2020).

Several NL2SQL approaches in this era focused on *schema understanding* and *schema linking* — the generalizability of PLMs required advanced techniques on ensuring that models both understand the semantic intent of users when querying and correctly identify database schema elements in natural language queries. Thus improving semantic accuracy of generated SQL queries. STRUG (Structure-Grounded-Pretraining) was introduced in 2020 by Deng et al. and presented a novel pretraining approach that improves model abilities when it comes to *schema linking*, it separates the problem in three facets: column grounding, value grounding and column-value mapping. In direct comparison with GRAPPA, STRUG achieves similar performance while being significantly cheaper to train (Deng et al., 2020).

In parallel, Zhong, Lewis, Wang, and Zettlemoyer released GAZP (Grounded Adaptation for Zero-shot Executable Semantic Parsing) in 2020. Zhong, Lewis, Wang, and Zettlemoyer specifically addressed the challenge of adapting semantic parsers across databases / domains which was a apparent problem with neural approaches which had a strong tendency to overfit on benchmark datasets. Its novel contribution was the combination of forward semantic parsing with a backward utterance generator which allowed for data synthesis in unseen environments which could then be used to adapt the semantic parser (Zhong et al., 2020). This approach enables a improvement in robustness and accuracy in situations where training and inference environments differ without requiring manually annotated examples (Zhong et al., 2020).

2.4.2 Advanced Pre-trained Language Model Approaches

Building on earlier foundational research on PLM application for NL2SQL tasks, researchers have developed increasingly complex systems that leveraged pre-trained language models whilst addressing their limitations when it comes to generating valid SQL.

Choi, Shin, Kim, and Shin introduced RYANSQL (Recursively Yielding Annotation Network for SQL) in 2020, which implements a sketch-based approach for decomposing complex SQL generation into multiple smaller problems. RYANSQL transformed nested statements into a set of top-level statements using the Statement Position Code (SPC) technique. This flattening of structure allowed RYANSQL to limit the complexity of the query generation problem whilst maintaining its ability to answer complex questions by recomposing complex queries from their parts. This approach allowed RYANSQL to achieve 58.2% accuracy on the SPIDER benchmark, representing a 3.2% improvement over contemporary state-of-the-art approaches at the time (Choi et al., 2020). The sketch-based approach makes RYANSQL a PLM-augumented successor of SQLNET which was a early neural approach to employ sketch-based query generation (Choi et al., 2020; Xu et al., 2017).

A significant advancement in terms of execution accuracy was reached with the application of T5-Models for NL2SQL tasks. T5 (Text-to-Text Transfer Transformer) Models have proven themselves as well-suited for for query generation — T5-3B for NL2SQL yielded 71.4% execution accuracy and thus presented a breakthrough in this domain of research (Rajkumar et al., 2022). This established a new baseline for PLM-based approaches and demonstrated that general-purpose language models could not only compete but outperform specialized architectures by far when properly fine-tuned (Rajkumar et al., 2022).

Following the adavancements through T5, J. Li, Hui, Cheng, et al. introduced GRAPHIX-T5 in 2023, which combined the T5 PLMs with a further graph-aware layers for NL2SQL tasks. This architecture could leverage both pre-trained knowledge of T5 models aswell as the database schema structure during inference. GRAPHIX-T5 constructs a schema graph where nodes represent tables and columns and edges represent relationships between them, such as foreign keys or columns association. This architecture allows the model to deeply

understand relationships and the layout of the database schema (J. Li, Hui, Cheng, et al., 2023). GRAPHIX-T5 outperformed standard T5 models significantly, with GRAPHIX-$T5_{LARGE}$ showing 6.6% increase in execution accuracy over $T5_{LARGE}$. When both GRAPHIX and the baseline T5 models were combined with PICARD (a novel constrained decoding mechanism) absolute δ between them jumped to 7.6% (81.0% in absolute numbers), evaluated on SPIDER-DEV (J. Li, Hui, Cheng, et al., 2023).

In parallel, H. Li, Zhang, Li, and Chen proposed RESDSQL in 2023, which proposed to decouple *schema linking* and *skeleton parsing*. This addressed the typical challenges sequence-to-sequence models faced when simultaneously trying to link both schema elements and generate the query skeleton (e.g. `SELECT <columns> FROM <table>`). RESDSQL further employed a ranking approach to filter schema elements before passing them to the model for query generation, which reduced noise (when working with large database schemas) and enabled passing only the most relevant parts. This twofold approach allowed RESDSQL to achieve state-of-the-art performance when being evaluated on SPIDER, outperforming GRAPHIX-$T5_{3B}$-PICARD by 0.8% in execution accuracy. When combined with NATSQL (a contemporary intermediate representation approach introduced by Gan et al. (2021)) absolute improvement over GRAPHIX-$T5_{3B}$-PICARD jumped to 3.1% emphasizing the robustness gain decoupeled architectures have over model-oriented approaches.

These advancement showed rapid improvements over earlier methods — far surpassing neural approaches — through advanced mechanisms when it comes to schema understanding and query generation. The wide language understanding inherited from PLM-basemodels further strengthens robustness and shows effectiveness through a large gain on the SPIDER benchmark. Collectively these approaches represent a leap in NL2SQL research, emphasizing their usability potential and real-world feasibility. This era primed the research field for the transition towards large language model adoption.

2.4.3 Constrained Decoding and Ranking Techniques

A major challenge in NL2SQL research is making sure that model generated queries are not just semantically accurate but also syntactically valid queries and thus executable. To address this issue Scholak, Schucher, and Bahdanau released PICARD (Parsing Incrementally for Constrained Auto-Regressive Decoding) in 2021, a constrained decoding mechanism for language models which utilizes the SQL grammar and constrained decoding mechanisms to incrementally parse the generate SQL, rejecting invalid tokens based on the grammar. PICARD showed to significantly improve the performance of pre-trained language models (like T5 or BERT) when it comes to NL2SQL tasks, lifting them from mid-level to state-of-the-art solutions on the SPIDER benchmark (Scholak et al., 2021).

PICARD operates as a incremental parser during model output decoding of pre-trained language models and continuously evaluates the probability of each token. Instead of just passing model outputs to a database for execution PICARD incrementally parses and validates the generated SQL, rejecting tokens if neeeded thus significantly improving the valid output accuracy (sometimes referred to as VA) of language models. This approach is addressing a significant issue associated with pre-trained language models — while they outperform in natural language understanding and reasoning, they often lack SQL grammer knowledge and tend to generate queries that are not executable due to ther unconstrained output space (Scholak et al., 2021).

The above introduces RESDSQL built ontop of PICARD's foundations and used a ranking-enhanced framework for input encoding. These two approaches represent a unqiue class of approaches that utilize input and output constraining in order to increase the performance characteristics of pre-trained language models (H. Li et al., 2023).

2.4.4 Advantages of PLM Approaches

PLM approaches to NL2SQL tasks have yielded significant performance improvements for the NL2SQL domain and represent a leap in NLIDB-research. They primed the research field towards using language models which led to a transition towards large language models in the following years. Namely PLM approaches brought a series of upsides with them:

1. **Compute Efficiency** — PLMs like RESDSQL achieve high accuracy (up to 84.1% on SPIDER depending on variants) wilst using far fewer parameters than conteporary LLMs, making them significantly more efficient and therefore reduce hardware requirements for their deployment (H. Li et al., 2023).

2. **Transferability** — Approaches like GRAPPA and STRUG can incorporate domain-specific understanding of natural language, table structures and SQL syntax during pre-training which addressed one of the primary issues with neural approaches (Deng et al., 2020; Yu et al., 2020).

3. **Vocabulary** — PLMs offer a larger vocabulary due to the vast amounts of training data available. This enables them to handle a wide variety of natural language patterns which addresses the benchmark-overfitting tendency of neural approaches which primarily trained on the development sets of contemporary benchmarks.

2.4.5 Limitations of PLM Approaches

Although representing the state-of-the-art at the time, PLMs introduce a class of problems which are associated with their non-NL2SQL associated nature. There have been an array of approaches to mitigate these shortcommings but nonetheless they must be considered when using a PLM-based approach to NL2SQL:

1. **Fine-tuning Requirements** — Most PLMs require substantial domain-specific, or at least NL2SQL specific, fine-tuning, limiting a straight forward adaptation to new domains or databases. Although being significantly more efficient than LLM-based approaches the potential need for initial fine-tuning represent a significant computational resource burden. Furthermore when not using synthetic data generation (e.g. Grappa) annotated datasets of training data are needed to achieve appropiate performance characteristics (Yu et al., 2020).

2. **Wide Input & Output Space** — Due to the general nature of PLMs their input and output space is often far larger than needed NL2SQL tasks. "Large pre-trained language models for textual data have an unconstrained output space; at each decoding step, they can produce any of 10,000s of sub-word tokens" (Scholak et al., 2021). This applies to both the input and output token space, therefore multiple approaches have been researched which focus on constraining these to the subset needed for NL2SQL tasks. Namely GRAPHIX-T5 and Picard have proposed potential (and promising) solutions to this issue (J. Li, Hui, Cheng, et al., 2023; Scholak et al., 2021).

3. **Limited Schema Awareness** — Due to being general purpose, and non-NL2SQL optimized, PLMs tend to incorporate limited amounts of schema awareness when being applied out of the box for NL2SQL tasks. Multiple research efforts focused on improving this situation, most notably Resdsql and GRAPHIX-T5 tried to improve the schema linking & awareness of PLMs (H. Li et al., 2023; J. Li, Hui, Cheng, et al., 2023), nonetheless the non-specialized nature of PLMs prevents NL2SQL being part of the fundamental model architecture.

These characteristics positioned PLMs as powerful but comparatively resource-intensive solutions for NL2SQL (especially in direct comparison with neural approaches), ultimately yielding the research domain to transition toward exploring Large Language Model approaches that promise even greater flexibility in adaptation and potentially superior handling of complex queries through advanced in-context learning approaches.

2.4.6 Comparison with Large Language Models

The research on applying pre-trained language models for NL2SQL tasks primed the field for the transition towards LLM usage. While PLMs like T5 and BERT range from millions to a few billion parameters, prevalent LLMs such as GPT-3 and GPT-4 operate at significantly larger scales, ranging from a few billions to hundred of billions parameters. The scale of LLMs enables in-context learning techniques that enable significantly easier and cheaper transferability of NL2SQL systems across domains (D. Gao et al., 2023). The δ of deployment, inference and training requirements of these two approaches are significant due to the size difference in models, which transfer to hardware requirements and therefore cost. While PLMs can require extensive fine-tuning on domain-specific data which may aswell be resource intensive (Deng et al., 2020; H. Li et al., 2023; J. Li, Hui, Cheng, et al., 2023; Yu et al., 2020), LLMs transfer the cost to the inference environment, where model modificants are less impactful, due to the extensive pre-training that took place. Approaches like DinSQL show that with the application of LLMs the engineering challenges around model instruction gained relevance while model training became less of a central problem to solve (Pourreza & Rafiei, 2023).

2.5 Large Language Models

The emergence of LLMs such as GPT-3, GPT-4, and Claude fundamentally transformed the landscape of NL2SQL research. Early experiments with LLMs for NL2SQL tasks showed state-of-the-art capabilities in comparsion with contemporary PLM approaches (D. Gao et al., 2023). Rajkumar et al. (2022) demonstrated

that CODEX (a contemporary model based on GPT-3), without any fine-tuning efforts, could achieve competitive performance on SPIDER, outperforming many state-of-the-art approaches that required extensive training. This breakthrough challenged the contemporary assumption that further specialization of model architectures would yield increases in NL2SQL performance (e.g. GRAPHIX-T5) (J. Li, Hui, Cheng, et al., 2023).

2.5.1 In-Context Learning

In-Context Learning (ICL) is a foundational approach for leveraging the ability of LLMs to utilize larger context windows for inference than traditional PLMs. Typical context windows of state-of-the-art LLMs can reach up to hundred thousands of tokens. This characteristic of LLMs enabled researchers to utilize this context window to provide examples of accurate NL2SQL translation instead of applying parameter updates. This paradigm shift has made developing NL2SQL systems significantly more accessible.

The fundamental principle of few-shot learning for NL2SQL involves providing the LLM with a small number of example pairs of natural language and their corresponding SQL representation. These examples can benefit the model's understanding of mapping between natural language and SQL syntax. This essentially builds on top of prior research like GRAPPA and STRUG, but applying these examples at inference time, rather than training time. Although this increases the inference cost of such a system, the upsides lie primarily in the flexibility of such an approach — database content / prior usage of the system can be dynamically utilized, rather than requiring retraining.

Example selection strategies showed to have a considerable impact on ICL performance. D. Gao et al. (2023) evaluated various example selection methods like *Random*, *Question Similarity Selection (QTS)*, *Masked Question Similarity Selection (MQS)*, and *Query Similarity Selection (QRS)*. D. Gao et al. propose a novel strategy to select, organize and present ICL examples to LLMs. DAIL-SQL utilizes both question and query similarity, masking domain-specific words and prioritizing examples that exceed a similarity threshold of τ (D. Gao et al., 2023, p. 5). DAIL-SQL encodes examples as question-SQL-pairs without the respective schema to improve token efficiency. Using a Code Representation Prompt (CR) for question and schema encoding yielded DAIL-SQL to achieve state-of-the-art 86.6% execution accuracy on SPIDER.

The comparison between zero-shot and few-shot performance reveals the accuracy gain potential through supplying examples to models in the inference context. While contemporary LLMs (such as GPT-4) have demonstrate impressive zero-shot performance (achieving 72.3% execution accuarcy on benchmarks like SPIDER) (D. Gao et al., 2023, Table 1, p. 8), few-shot learning still shows to substantially improve model performance. D. Gao et al. (2023) shows that even one-shot learning boosts GPT-4's execution accuracy to 80.2%, representing a 7.9% increase, while five-shot learning reaches 82.4% (**?**, Table 2, p. 8).

Especially with complex queryies which can involve multiple tables, nested queries and complex joins, zero-shot approaches often dramatically underperform k-shot ones. NL2SQL approaches that dont supply examples to the model during inference time fail more frequently to generate semantically accurate SQL queries (D. Gao et al., 2023). Notable is the leap in exact match ratio measured by the SPIDER benchmark — jumping from 22.1% for GPT-4 using zero-shot to 71.9% with five-shot. The results presented by D. Gao et al. (2023) show signficant correlation between k and the execution accuracy of k-shot approaches.

This effectiveness has established ICL approaches as a standard technique applied in LLM-based NL2SQL approaches. Contemporary approaches like XIYAN-SQL, CHASE-SQL AND DIN-SQL all utilize variations of ICL to achieve state-of-the-art results (Y. Gao et al., 2025; Pourreza et al., 2024; Pourreza & Rafiei, 2023).

2.5.2 Self-Correction and Iterative Refinement

Pourreza and Rafiei (2023) proposed DIN-SQL as an innovative approach to NLIDBs that rely on LLMs. DIN-SQL decomposes complex queries into sub-parts and utilizes in-context-learning and self-correction during the generation phase. Compared to DAIL-SQL which relies on example selection during the in-context-learning phase, DIN-SQL focuses on a refinement loop that allows the model to self- correct errors it made during the initial generation phase — thus the model can repair schema linking, syntactic or semantic errors. By explicitly instructing the LLM to review its work against a specfic schema, the user input and potential database errors, DIN-SQL achieves a high execution accuracy on SPIDER with 85.3%. Therefore DIN-SQL outperforms contemporary approaches but is surpassed by by DAIL-SQL by 1.3% (D. Gao et al., 2023; Pourreza & Rafiei, 2023). Furthermore DIN-SQL makes observations on the impact that the self-correction prompt can steer results significantly — Pourreza and Rafiei found that using *generic self-correction* (ie. assuming the query contains errors) lowers the execution accuracy by 4.2% on SPIDER compared to *gentle self-correction* (ie. assuming nothing about the validity of the query). It was noted that the impact of the self-correction mechanism relies on

the model size, with smaller models performing better with *generic self-correction* and larger models performing better with *gentle self-correction* (Pourreza & Rafiei, 2023). The self-correcting nature of DIN-SQL represents a diversion from DAIL-SQL's emphasis on input optimization towards output refinement. Pourreza and Rafiei demonstrate how structured introspection can play a significant role in enhancing LLM performance for formal language generation tasks.

Building upon DIN-SQL's self-correction module, Askari, Poelitz, and Tang (2024) proposed MAGIC (Multi-Agent Guideline for In-Context Text-to-SQL), which further advances the self-correction mechanism through harnessing a set of specialized agents to automate the self-correction prompt engineering (Askari et al., 2024). MAGIC consists of a manager agent, a correction agent and a feedback agent that collaboratively refine LLM instructions during the refinement loop. Further MAGIC derives common faliure patterns of the initial query generation phase from training data, allowing it to efficiently spot the most common mistakes that the model makes at generation time. This approach represent a further advancement on Pourreza and Rafiei's DIN-SQL, effectively supersetting the *generic* and *gentle* correction mechanisms through an intelligent, self-adapting one (Askari et al., 2024). The autogenerated guidelines from MAGIC yield 85.6% execution accuracy on the Spider development set — representing a 5.31% improvement over DIN-SQL's human written correction guidelines. These results emphasize that optimized self-correction mechanisms have the ability to significantly drive up overall system performance of NLIDBs (Askari et al., 2024).

While DIN-SQL and MAGIC focus on automated self-correction in single-turn settings, H. Zhang, Cao, Xu, Chen, and Yu (2024) introduced the concept of Chain-of-Editions (CoE-SQL), which addresses the unique challenges of multi-turn conversational NLIDBs. Conversational interfaces for NL2SQL systems enable human-in-the-loop refinement. Interactive information seeking from the user has shown to be an effective way to drive overall accuracy of the system and improve user satisfaction (F. Li & Jagadish, 2014). Rather than approaching each query independently, CoE-SQL recognizes that in a conversational context, successive SQL queries usuallly require only small and incremental modifications of the previous queries. Interactive user input is an effective measure for dealing with ambiguous natural language queries (F. Li & Jagadish, 2014; Montgomery, 1972; H. Zhang et al., 2024).

2.5.3 Candidate Selection Frameworks

Contemporary NL2SQL approaches have increasingly emphasized on the generation of query candidates and their selectiion as a promising architecture. Candidate selection strategies have shown significant performance improvements on challenging benchmarks. These approach acknowledge the inherent difficulty of generating perfect SQL queries in one attempt / using one generation mechanism, even with capable LLMs and modern self-correction mechanisms.

Pourreza et al. (2024) introduced CHASE-SQL in 2024, a framework that leverages multiple reasoning paths to generate multiple query candidates. After the initial generation phase CHASE selects the most promising solution to the natural query input. CHASE-SQL harnesses three different generation strategies: A divide-and-conquer approach which breaks down complex natural language queries into multiple sub tasks that can be individually tackled, a chain-of-thought based generation approach which inspects execution plans of SQL queries and a schema-aware generation of synthetic examples that can be used for in-context learning (Pourreza et al., 2024). These different generation mechanisms produce a set of query candidates that each have different characteristics. For candidate selection CHASE harnesses a fine-tuned LLM that can do binary selection of candidates. Pourreza et al. have demostrated that their query selection approach is more robust than apparent alternatives and yields state-of-the-art performance with 73% execution accuracy on Bird and 87.6% on Spider (Pourreza et al., 2024).

Conceptually similar work has been done by Y. Gao et al. with XiYan-SQL which is architected as multi-generation ensemble strategy with better schema representation. XiYan-SQL integrated in-context learning alongside supervised fine-tuning approaches to generate query candidates (Y. Gao et al., 2025). A key contribution of Y. Gao et al. (2025) is their M-Schema representation of database schemas, which improves the models schema awareness and reduces frequent schema linking errors. XiYan-SQL enhances accuracy by utilizing multiple different strategies that have complementary characteristics during query generation to enhance the robustness of the overall system. The query generation stage utilizes both a fine-tuned SQL generation model aswell as ICL strategies to achieve a breadth of candidate coverage. Following to the query generation stage a self-correction stage (referred to as *refinement* stage by Y. Gao et al.) is utilized to correct common errors. Lastly a selection model is used to choose the most accurate candidate that was produced during the generation stage (Y. Gao et al., 2025). Through this sophisticated and diverse architecture XiYan-SQL was able to

achieve impressive results across contemporary NL2SQL benchmarks — achieving 89.65% execution accuracy on Spider and 73.34% on Bird, which renders XiYan-SQL state-of-the-art (Y. Gao et al., 2025).

Both CHASE and XiYan-SQL show that diversifing candidate generation and training specialized models for candidate selection yield state-of-the-art execution accuracy which significantly outperforms single-path generation approaches. The success of these two approaches indicates that for increasingly complex NL2SQL tasks (such as Spider2.0), the capability to generate multiple valid interpretations of the natural language query is an important stepping stone to achieving meaningful execution accuracy. Both CHASE and XiYan-SQL rely on specialized candidate selection models which renders these approaches to combine the strengths of LLMs when it comes to language understanding and transferability with the robustness of specialized model architectures for candidate ranking and selection.

2.5.4 Retrieval-Augumented Generation

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for enhancing NL2SQL systems by integrating external knowledge retrieval with the generative capabilities of LLMs. This technique is seen in all above introduced papers in varying forms (Askari et al., 2024; D. Gao et al., 2023; Y. Gao et al., 2025; Pourreza et al., 2024; Pourreza & Rafiei, 2023; H. Zhang et al., 2024). The most prevalent form of RAG in NL2SQL is the encoding of the database schema into the LLM prompt for in-context-learning. This allows LLMs to be aware of the table structures and names, foreign key relationships, primary keys etc. Y. Gao et al. (2025) proposed the serialization of database schemas in the M-Schema format, a semi structured, text-based schema description. D. Gao et al. (2023) proposed to encode the schema using a Code Representation Prompt (CR) which refers to the encoding of the raw SQL statements need to construct the schema. Y. Gao et al. (2025) provided an ablation study for the M-Schema which yielded questionable results on the optimization of this approach. While the M-Schema format yielded best results when XiYan-SQL was combined with GPT-4o or Claude 3.5 Sonnet, it performed worse than alternatives on DeepSeek and Gemini models (Y. Gao et al., 2025).

A relevant optimization technique for RAG is the selection of a schema subset before encoding the schema for the model. ResdSQL was one of the earlier approaches to explore subset-encoding of database schemas with Shen et al. (2024) building on top of this (H. Li et al., 2023; Shen et al., 2024). Shen et al. (2024) introduced ASTReS which dynamically retrieves database schemas and uses abstract syntax trees (ASTs) to select optimal few-shot examples for ICL. By pruning the ASTs down to the most relevant subset, ASTReS achieved the highest at-the-time (2024) execution accuracy on Spider with 86.6% indicating that subset-encoding is a sensible optimization mechanism. ASTReS was combined with GRAPHIX-T5 in order to achieve this result (Shen et al., 2024).

The impact of RAG when NL2SQL systems face large and complex database schemas has been particularly significant. Traditional approaches struggle when database schemas contain hundreds or thousands of tables and columns, as the complete schema may not fit within model context windows depending on their size. RAG-based systems address this by dynamically retrieving only the most relevant portions of the schema based on the natural language query. The technique of subset-encoding becomes especially relevant in enterprise environments where database schemas can be extremely large and complex. Recent benchmarks like Spider2.0 emphasize enterprise environments and show that existing solutions underperform in those scenarios, often reaching single digit execution accuracy.

The work of Shen et al. indicates that prefiltering of the environment of language models is an effective and promising technique that has the ability to reduce computational requirements of contemporary NL2SQL systems. As introduced above, recent state-of-the-art systems often utilize closed-source models like Gemini, GPT-4(o), Claude 3.5/3.7 Sonnet etc which often come with massive parameter sizes (reaching hundreds of billions of parameters). ASTReS demonstrated that efficient schema retrieval mechanisms enable smaller models (e.g. GRAPHIX-T5) to achieve state-of-the-art performance against LLM based approaches like DAIL-SQL (D. Gao et al., 2023; Shen et al., 2024).

2.5.5 Specialized LLMs and Fine-tuning

While general-purpose LLMs demonstrated state-of-the-art natural language understanding capabilities, the research domain of NLIDBs increasingly explored the potential of fine-tuned LLMs for NL2SQL tasks, which offer a promising tradeof between natural language understanding (ie. breadth of the model) and concrete SQL generation capabilities (ie. depth of the model). Multiple research works have been done on the fine-tuning of language models which yielded a series of dedicated models for optimized SQL generation.

A significant limitation of many LLM-based NL2SQL solutions is their dependency on proprietary and closed-source LLMs like GPT-4(o), Claude and Gemini. Whilst they are useful for initially proving the potential of LLMs on contemporary benchmarks, this dependency introduces significant concerns related to data-privacy, high-side use (e.g. in classified environments), transparency of data-flow and deployment costs. To address these challenges H. Li et al. (2024) have introduced CodeS in 2024, a series of of open-source language models dedicated for NL2SQL tasks with parameter sizes ranging from 1B to 15B. The CodeS models were evaluated against contemporary benchmarks like Spider and Bird and showed promising inference results when compared to their closed-source counterparts. H. Li et al. (2024) showed that CodeS 7B achieves 85.4% execution accuracy on the Spider development set, outperforming both fine-tuning approaches like RESDSQL and GRAPHIX-T5-PICARD aswell as prompting based methods DIN-SQL+GPT-4 and DAIL-SQL+GPT-4 (H. Li et al., 2024). The same tendency was observed on Bird with CodeS-15B achieving 60.37% execution accuracy, compared to 57.41% for DAIL-SQL+GPT-4 and 55.90% for DIN-SQL+GPT4 (H. Li et al., 2024). This marks a significant advancement in the open-source language model research area, with CodeS reaching new state-of-the-art performance in 2024. While more recent approaches like XiYan-SQL and CHASE both outperform CodeS, CHASE relies on proprietary models and XiYan-SQL doesnt provide any information on what base models where used.

H. Li et al. (2024) addressed serveral critical research challenges in the NL2SQL domain and proved that open-source models could perform competitively with proprietary models whilst maintaining a significantly smaller parameter footprint (ie. 7B and 15B) compared to GPT-4 which is a multi-hundred-billion parameter model. This makes it feasibile to deploy CodeS locally, instead of relying on an enterprise API like OpenAI's one (H. Li et al., 2024), therefore making it highly practical for real-world deployments where computation resource are constrained.

H. Li et al. published a follow-up paper in 2025 which introduced the next-generation OmniSQL models with 7B, 14B and 32B sizes, trained using synthetic data generation. OmniSQL achieves state-of-the-art performance when compared to alternative LLMs - including both open-source and closed-source competitors. The models achieve significant execution accuracy improvements on both Spider and Bird, the 7B model reaches 88.9% on Spider and 66.1% execution accuracy on Bird which represents a significant (5%+) improvement over comparable alternatives (H. Li et al., 2025). These results were achieved without combining OmniSQL with advanced in-context-learning, self-correction, retrieval-augumented-generation or candidate-selection techniques, which indicates that even higher accuracy scores are possible.

2.5.6 Limitations and Challenges

Despite the impressive advancements of LLM-based approaches for NL2SQL, several significant limitations and challenges are apparent that impact their practical implementation and real-world deployability.

1. **Hallucination and Accuracy Concerns** — Since LLMs demonstrate a tendency to hallucinate, LLM-based NL2SQL systems face the challenge of detecting when LLMs generate plausible but incorrect SQL queries. The breadth of possible errors ranges from detectable errors like invalid table or columns references, invalid SQL syntax etc. to undetectable, slight errors in semantics, like reversing the order of aggregations or using the wrong aggregation method. Contrary to traditional approaches which fail visibly with unknown structures, LLM-based approaches might produce semantically flawed queries that execute without errors and yet return semantically incorrect data. As noted by Floratou et al., "achieving Enterprise-Grade NL2SQL is still far from being resolved" even with state-of-the-art models, particularly when handling complex real-world database schemas and ambiguous queries (Floratou et al., 2024; Lei et al., 2025).

2. **Computational Resource Requriements** — The resource intensity of LLM-based systems presents barriers to widespread adoption. High-performance state-of-the-art models require substantial computational resources during inference, making them expensive to deploy at scale. While smaller models exist, they typically show reduced performance on complex queries, creating a challenging trade-off between accuracy and resource efficiency. Studies like DAIL-SQL demonstrate a direct correlation between model size and performance, where the most accurate systems are also the most challenging to deploy economically (D. Gao et al., 2023). Even state-of-the-art specialized LLMs like OmniSQL degrade in performance as parameter sizes shrink, although they maintain a significantly higher parameter efficiency, outperforming enterprise-level closed-source competitors.

3. **Data Privacy and Security Implications** — Using LLMs particularly through proprietary APIs brings along considerable concerns with regards to data privacy and security as sensitive data may be transferred to respective model vendors. Potential senstive data aswell as database schemas alongside user questions needs to be communicated to external parties in order to form a usable system if closed source models are used. Recent open-source model developments like CodeS and OmniSQL seem to mitigate this problem partially, but as mentioned above using local inference with LLMs brings along stark computational requirements.

4. **Ambiguity Handling in Complex Scenarios** — Following previous paradigms LLMs also continue to struggle with effective ambiguity resolution in natural language queries over complex database schemas. These challenges are particularly prominent in large scale enterprise environments where similar column names exist across multiple tables, or domain-specific terminology has multiple potential interpretations that can result in different queries. Though approaches like multi-path reasoning and candidate selection show promising improvements in execution accuracy, they often increase inference time and complexity as they run multiple parallel inference steps and rank their results (Y. Gao et al., 2025; Pourreza et al., 2024).

5. **Competitive PLM Approaches** — Prevalent PLM-based approaches like RESDSQL and GRAPHIX-T5 achieved impressive execution accuracy while using significantly fewer parameters than LLMs, making them more computationally efficient (H. Li et al., 2023). Despite reducing inference-time requirements, these approaches typically utilize re-training and domain specific fine-tuning which both limits their flexibility and introduces training costs. Yet for certain scenarios PLM-based approaches offer an interesting tradeoffs: Frontloading the computational effort can be interesting in compute-constrained environments and help to reduce ongoing costs. LLMs dont require the upfront cost of fine-tuning and adaption but require significantly more resource for their deployment. This indicates that certain systems and environments might benefit more from PLM approaches than LLM ones. Contemporary research, which typically focuses around state-of-the-art-performance on benchmarks, shifted primarily towards LLMs due to their transferability and natural language coverage.

2.6 Benchmarking

In order to evaluate NL2SQL systems standardized benchmarks like Spider and Bird have emerged. These benchmarks can measure performance across different approaches and models, enable meaningful ablation studies and are a useful indicator for the state of the research field. In the past decade significant advancements have been made with Spider being released in 2018 the first major, widely adopted, benchmark emerged in this field (Yu, Zhang, et al., 2018).

2.6.1 Spider

Spider, introduced by Yu, Zhang, et al. in 2018, has become the de facto standard benchmark for evaluating complex and cross-domain Text-to-SQL systems. It consists of 10,181 questions and 5,693 unique SQL queries spanning 200 databases across 138 domains. Previous benchmarks like lacked complexity and cross-domain distrubtion of datapoints which prevented the *transferability* of approaches or models to be accounted for in benchmarks. With Spider the capability to be database agnostic was required to achieve meaningful accuracy scores. Furthermore Spider was split in training and test sets which contain different database in order to prevent overfitting models from succeeding. This design specifically tests a model's ability to handle schema linking and generalization challenges rather than memorizing specific database patterns. Spider evaluates both *exact matching accuracy* and *execution accuracy*, with contemporary state-of-the-art systems achieving approximately 85-90% *execution accuracy* (as of 2025) (Y. Gao et al., 2025; H. Li et al., 2025; Pourreza et al., 2024; Yu, Zhang, et al., 2018).

2.6.2 Bird

The Bird benchmark (BIg bench for large-scale database gRounded Text-to-SQLs), released in 2023, and addresses the gap between academic benchmarks and real-world applications by focusing on large-scale databases with actual data content (J. Li, Hui, Qu, et al., 2023). Bird contains 12,751 text-to-SQL pairs and 95 databases with a total size of 33.4 GB across 37 professional domains. Unlike Spider, which primarily evaluates against database schemas with minimal content, Bird emphasizes challenges related to dirty database contents, external

knowledge between natural language questions and database values, and SQL efficiency in massive databases. This places BIRD as a relevant benchmark for real world feasibility of approaches and models. Even state-of-the-art LLMs like GPT-4 achieve only 54.89% execution accuracy on BIRD, compared to human performance of 92.96%, highlighting the significant challenges posed by real-world database scenarios on NLIDBs (J. Li, Hui, Qu, et al., 2023).

2.6.3 Spider 2.0

SPIDER 2.0 which was introduced by Lei et al. in 2025 represents the most recent advancements of benchmarks for NL2SQL systems. It represents a significant evolution in NL2SQL benchmarking and focuses primarily on enterprise level database challenges. SPIDER 2.0 is much smaller with only 632 real-world problems which were derived from enterprise database usecases, but yet represents a meaningful indicator for the complexity of databases that approaches and models can handle. SPIDER 2.0 goes beyond simple query generation tasks and moves towards testing the deep understanding of the database, requiring models to understand metadata, SQL dialect documentation and project-level codebases (Lei et al., 2025). The tasks contained in SPIDER 2.0 often demand multiple complex SQL queries often exceeding the 100-line mark and require incorporating a diverse set of database operation from transformation to analytics. Lei et al. further highlights the gap between academic research and enterprise-level environments with even advanced approaches achieving only 21.3% on SPIDER2.0 compared to 91.2% on SPIDER 1.0 (Lei et al., 2025; Yu, Zhang, et al., 2018).

2.7 Research Gaps

This literature review highlights that significant progress was made in the development of real world feasible NL2SQL systems and that the research domain has undergone multiple large paradigm shifts from rule-based to neural-network-based to PLM-based and most recently to LLM-based approaches. Despite these advancement several critical research gaps remained open or emerged which limit the practical deployment of NL2SQL systems and their widespread adoption as natural language database interfaces in real world scenarios.

2.7.1 Advanced Open-Source Approaches

While recent research papers have made remarkable progress on both open-source model development (CODES and OMNISQL) aswell as advanced architecture development like CHASE and XIYAN-SQL (which utilize multi-path generation and candidate selection, self-correction and RAG), a significant research gap evolved around state-of-the-art approaches that dont rely on proprietary LLMs and only utilize open-source models. Most contemporary research proposes solutions which (partially) rely on the baseline proprietary LLM in order to achieve state-of-the-art results. Whilst these efforts give meaningful signal to the effect that specific prompt engineering techniques, self-correction mechanisms and candidate selection models have, they are not yet feasible for real world scenarios due to the cost and data privacy concerns they introduce.

1. **Limited Exploration of Technique Synergies** — Contemporary research primarily focuses on techniques like candidate selection, RAG, and self-correction in isolation or with specific closed-source LLMs. A gap remains in understanding how these techniques can be optimally combined, especially with emerging open-source models. For instance, the potential of synergy of specialized open-source LLM like OMNISQL with candidate-selection, advanced self-correction and subset-encoding of database schemas has yet to be researched.

2. **Efficiency-Accuracy Tradeoffs** — Whilst recent research efforts focus primarily on achieving a new state-of-the-art execution accuracy metric on prevalent benchmarks, the relationship of computational requirements and performance gains achieved remains unclear. Some research papers present ablation studies indicating the relevance of certain architecture components but the overall relationship between state-of-the-art performance and cost is still underexplored.

3. **Cross-Technique Optimization** — There is limited research on how to optimize the interplay between the various NL2SQL techniques. For example, it is yet unclear how subset-encoding of database schemas might impact the effectiveness of multi-path generation and candidate selection, or whether synthetic example generation for in-context-learning is effective when working with specialized models versus general-purpose LLMs.

2.7.2 Deployment and Performance Gaps

Despite impressive academic results, significant gaps remain in transitioning NL2SQL systems from research environments to real-world, enterprise-feasible, deployments:

1. **Database Integration** — While existing research has often focused on standalone systems, little attention has been given to the integration complexities between NL2SQL capabilities and database management systems like PostgreSQL. Having two standalone systems imposes a significant data transfer need between the two systems when infact, natural language queries can be treated as an extension to most existing databases. This implementation gap prevents seamless adoption into existing databases as it requires additional software layers which in turn increase the overall complexity and cost.

2. **Hardware Requirement Optimization** — As mentioned above, current LLM-based approaches often require significant compute resources in order to achieve state-of-the-art results. There is limited research available on performance optimization of NLIDBs in order to achieve practical and industry-viable hardware constraints.

3. **Real-time Performance Considerations** — Most research papers evaluate models based on their accuracy metrics alone without factoring in the latency and throughput characteristics of their solutions. This imposes a possibly significant δ between academic research and production environments. Responsiveness is an important metric for user experience and should therefore play a role in NLIDB research when it comes to evaluating different NL2SQL architectures.

4. **Privacy and Security** — While open-source models address some privacy concerns by enabling a local deployment of LLMs, research gaps remain in ensuring that NL2SQL systems respect database access controls and security policies. This is especially important for enterprise and government environments where data access is strictly regulated.

2.7.3 Ambiguity Resolution and Semantic Accuracy

Despite the long existance of the research field, fundamental questions remain open with regards to handling ambiguity effectively in natural language queries. Current, especially language model based, systems struggle to correctly identify when natural language queries contain ambiguities that they cant confidently resolve. Although some approaches like multi-path generation and candidate selection are a promising way to work around ambiguous language, early systems like NaLIR showed that the most effective way to deal with inherently ambiguous language is asking clarifying questions, instead of trying to interpret the query best-effort (F. Li & Jagadish, 2014). Ambiguous language is an inherent source of inaccuracy and therefore a cause for misleading query results. Contemporary research and benchmarks dont focus ambiguity detection and strategies resolution which therefore leaves open questions to be further researched.

2.7.4 Evaluation and Benchmarking Gaps

1. **Enterprise-grade Evaluation** — As highlighted by Spider 2.0, there is a gap between academic benchmarks and enterprise realities. Further research is needed to create evaluation frameworks that better represent real-world enterprise environments with thousands of tables and complex relationships. Lei et al. (2025) is a promising first step is this direction.

2. **Performance Metrics** — As mentioned above, current benchmarks often dont capture latency or throughput of NL2SQL systems at all, allowing for solutions to achieve state-of-the-art scores that require significantly more resources than their predecessors. Having meaningful performance metrics and benchmarking would allow to further analyze the proposed solutions for their real world feasibility.

3. **User Experience Metrics** — Contemporary benchmarks focus on execution accuracy without assessing potential user satisfaction, trust, and overall experience. Metrics that capture these aspects are necessary for understanding the actual utility of NL2SQL systems in practice. Although execution accuracy is a big factor for the trustworthiness of a NLIDB, it is not the only component as indicated by F. Li and Jagadish in 2014.

2.7.5 Thesis Placement

This thesis addresses several of the above outlined research gaps. The primary focus of this work is the integration of open-source NL2SQL models with advanced techniques like candidate selection, subset-encoding of database schemas, and synthetic example generation for in-context learning. Through the implementation of a PostgreSQL extension this work will bridge multiple critical gaps between theoretical advancements and their practical deployability into real world systems whilst exploring performance characteristics.

3 Theoretical Foundations

This section introduces the theoretical concepts needed for understanding NATURAL's design and implementation sections. It covers relational database theory, machine learning fundamentals, large language models, natural language processing, embeddings and graph theory.

These foundations directly support the system's core functions: example selection (σ), schema subsetting (ϕ), query projection (π), self-refinement (ρ), and voting (ν).

3.1 Relational Database Theory

The relational database theory provides a mathematical framework for formalizing data layout in database management systems (DBMS), enabling efficient storage and retrieval mechanisms. This theory is helps to understand the way in which natural language queries need to be translated into database operations and what challenges arise during this translation.

3.1.1 Relational Model Fundamentals

The relational model was introduced by Codd in 1970 and essentially forms the theoretical foundation for most modern database systems (Codd, 1970, 1974). A relational database organizes data in relations (ie. tables), where each relation consists of tuples (ie. rows) containing attributes (ie. columns) of specific domains (ie. data types and constraints). The mathematical origins of this model root in set theory and first-order predicate logic (Codd, 1970; Date, 2003).

Furthermore the relational model abstracts data storage and respective physical implementation details by design, allowing interactions with data through respective logical structures (relations, tuples, attributes etc.) rather than exposing the underlying storage mechanism. This abstraction is fundamental for NL2SQL systems, as natural language queries can be mapped into queries on these structures regardless of underlying physical storage implementation, allowing for better transferability of NL2SQL systems (e.g. across different database deployments).

3.1.2 Core Concepts

The core concepts of the relational model that are relevant for this thesis include:

- **Relations** — Mathematical sets of tuples representing entities (e.g. employees, products) or relationships (e.g. enrollment, purchases). Each relation has a fixed structure.
- **Tuples** — Tuples represent rows of data, by grouping a several attributes into one logical unit (e.g. one specific purchase, containing the amount, date and shop). All tuples in the same relation share the same structure.
- **Attributes** — Attributes are frequently referred to as columns, representing a specific dimension of a tuple (e.g. customers have names). Attributes are represented in a fixed domain.
- **Domains** — Domains are sets of allowed values for attributes, including mathematically formalizing data types and constraints.
- **Schemas** — Structural definitions specifying relations, their attribute names, domains, and integrity constraints.
- **Primary Keys** — Attributes that uniquely identify tuples within a relation. These ensure entity integrity and ensure referential relationships.
- **Foreign Keys** — Attributes referencing primary keys in other relations, allowing databases to maintain referential integrity and enabling semantically correct complex queries across multiple tables through joins.

These concepts form the semantic foundation that natural language interfaces must navigate when translating user queries into formal database operations. As shown in the literature review *schema-linking* is a crucial problem of NLIDBs — establishing an understanding of which relations a natural language query refers to is the first step of formalizing a database operation for retrieval.

3.1.3 SQL as a Declarative Query Language

Structured Query Language (SQL) is a defacto standard for interfacing with relational databases. It is a declarative language to describe database operations like selection, insertion, updating and deletion of database conte/home/mara/Workspace/thb/thesis/main.texnts. SQL's design is strongly influenced by relational algebra and tuple relational calculus. It enables complex data retrieval through a readable query syntax incorporating relational algebra operations like:

- **Selection** (σ) — Filtering tuples based on specified conditions
- **Projection** (π) — Extracting specific attributes from relations
- **Joins** ($\bowtie$) — Combining data from multiple relations based on common attributes
- **Aggregation** — Computing summary statistics over groups of tuples

The basic syntax of SQL queries follows a logical pattern that reflects these theoretical operations:

```
SELECT attributes      -- Project attributes
FROM relations         -- Specify relations to use
WHERE conditions       -- Selection of tuples
GROUP BY attributes    -- Grouping for aggregation
HAVING conditions      -- Group-level filtering
ORDER BY attributes    -- Ordering of results
```

There are further operations like various types of `JOIN`s, `OFFSET`, `LIMIT`, `WITH` that are heavily used in SQL which are excluded for readability and simplicity.

The nature of SQL creates a fundamental challenge for NL2SQL systems; natural language queries express (often fuzzy) user intent in terms of desired output (e.g., "Give me 5 great movies") while SQL formalizes an explicit and discrete way of how to retrieve this information from the database:

```
SELECT m.title, AVG(r.score) as rating
FROM movies m
JOIN ratings r ON m.movie_id = r.movie_id
GROUP BY m.movie_id, m.title
HAVING AVG(r.score) >= 8
ORDER BY rating DESC
LIMIT 5
```

This semantic gap between natural language and formal queries represents the core challenge that this thesis addresses.

Furthermore, SQL's compositional nature allows complex queries to be built from simpler components through nesting and combination of operations. NL2SQL systems must therefore not only fill in the individual query components but also how to compose these to semantically accurate and syntactically correct query statements. This is particularly important when dealing with complex natural language queries that may require multi-clause SQL statements involving subqueries, multiple joins, and aggregation functions to answer.

3.1.4 Normalization and Schema Design

Database schemas typically follow normalization principles to eliminate redundancy and maintain data integrity. The normal forms (1NF, 2NF, 3NF, BCNF) provide a set of design guidelines and rules for relational database design to reduce duplication, inconsistencies and anomalies (Date, 2003). Understanding normalization forms is crucial for NL2SQL systems to handle:

- **Schema complexity** — Normalized schemas often distribute logically related information across multiple tables, requiring natural language interfaces to understand implicit relationships and generate appropriate joins.
- **Semantic mapping** — Humans typically think about data in denormalized, conceptual terms, while relational databases store data in its normalized form. NL2SQL systems must overcome this layout difference (e.g. there is no 1:1 mapping of concepts to tables).

- **Query complexity** — Retrieving simple information from relational databases may require multiple joins in normalized schemas, challenging NL2SQL systems to generate potentially complex SQL statements from simple user requests.

The tension between normalized database design and intuitive natural language use represents a key challenge that influences the architecture and design decisions explored in subsequent sections of this thesis.

3.2 Machine Learning Fundamentals

Machine learning represents the algorithmic foundation for state-of-the-art NL2SQL systems, enabling systems to learn how natural language queries translate into SQL. Understanding these foundations is essential for implementing and optimizing large language model-based NL2SQL approaches.

3.2.1 Neural Network Architecture

Neural networks are the computational concept behind contemporary NL2SQL approaches, representing complex capabilities through compositions of simpler operations. A neural network is comprised of interconnected computational units (called neurons) organized in layers, where each connection has an associated (and learnable) weight and bias parameters.

The fundamental computation paradigm of neural network outputs is called forward propagation, which is applying the respective weights (W) and biases (b) to the input parameter (a) and transforming it using a (non-linear) activation function f:

$$a^{l+1} = f(Wa^l + b) \tag{1}$$

where $W \in \mathbb{R}^{m \times n}$ is the weight matrix, $a \in \mathbb{R}^n$ is the input vector, $b \in \mathbb{R}^m$ is the bias vector, and f is an activiation function (e.g. ReLU, sigmoid, or tanh). Using this function repeatedly, propagates information through the neural network.

The most important architectures for simple natural language processing networks include:

- **Feedforward Networks** — Which can process fixed-size inputs through successive linear transformations and activations, this architecture is particularly suitable for classification and regression problems within NL2SQL systems.
- **Recurrent Networks (RNNs/LSTMs)** — Which can handle sequential data of variable-length by maintaining hidden states that capture dependencies, enabling processing of natural language sequences of arbitrary length.
- **Embedding Layers** — Which map discrete tokens (e.g. words, characters, or subwords) to a dense vector representation, thus providing the foundation for subsequent neural language processing.

The ability of neural networks to learn hierarchical representations through multiple layers makes them particularly well-suited for the complex translation required in NL2SQL systems.

3.2.2 Learning Approaches

Broadly speaking, there are three fundamental approaches to training a machine learning model, also known as the process of "learning":

1. **Supervised Learning** – The process of training a model using a labeled dataset of input and output pairs. This learning approach is quite commonly used in models used for NL2SQL which were trained on question and answer pairs (natural language question and corresponding SQL query). This learning approach aims to minimise the loss between predicitions and ground truth.
2. **Unsupervised Learning** – Includes learning patterns that work with unlabeled datasets and use approaches like clustering, language modeling and heuristics to infer structure and yield predictions. Unsupervised learning is most appropriate when there is no clear, singular answer and the domain provides large datasets that are not feasible to label manually (eg, dataset gathered through webscraping the internet).

3. **Reinforcement Learning** – The process of learning through reward signals, commonly known as the reward function. Applied to NL2SQL systems this could transfer to training a model's SQL generation capabilities based on execution feedback. Historically this learning pattern is not widely used in NL2SQL approaches.

3.2.3 Attention Mechanisms

Attention mechanisms allow machine learning models to selectively focus on relevant parts of their input by learning to weight different sections of the input sequence based on its relevance to the current computation. While RNNs processed sequences sequentially, attention enables direct modeling of relationships between any positions in a sequence, which enabled the transformer architecture to become prevalent for natural language processing (**?**).

Attention can be computed by retrieving the weighted sum of values V based on the compatibility between queries Q and keys K:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where d_k is the dimensionality of the key vectors, used for scaling to prevent softmax saturation which would incurr small gradients. The softmax operation converts compatibility scores into a probability distribution, determining how much attention to pay to each value.

This concept proved to be a foundational advancement to transformer architectures seen in natural language processing and subsequently recent NL2SQL research. Transformers extended the concept of attention into self-attention and multi-head attention patterns that underpin todays large language models.

3.2.4 Transformer Architecture

The transformer architecture, introduced by **?** in **?**, introduced the model architecture found in modern large language models (**?**). Unlike earlier sequence to sequence models, transformers process are capable of processing entire sequences of inputs in parallel by using a self-attention mechanism. This enabled more efficient training on massive datasets and superior representation of of complex dependencies in input sequences.

3.2.4.1 Self-Attention Mechanism

Self-attention allows each position in a sequence to be weighted to all other positions in the sequence, thus resulting in a representations that can capture contextual relationships between input tokens. For an input sequence of tokens $x_1, \ldots, x_n$, the self-attention mechanism computes:

$$\text{SelfAttention}(X) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ are learned linear projections of X into the query, key and value representations. This mechanism enables transformers to weigh the importance of different tokens while encoding each position. This maintains semantic and syntactic relationships which are crucial for understanding natural language queries in NL2SQL systems.

3.2.4.2 Multi-Head Attention

Multi-head attention extended the self-attention concept by computing multiple attention operations in parallel, allowing the model to put focus to different aspects of the input at the same time:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{4}$$

where each *head* computes:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

Each attention head can specialize in capturing a different type of semantic relationship. For example one head might focus on syntactic structure (e.g., linking SQL keywords to their arguments), while others might capture semantic relationships between works (e.g., List all x where y). This multi-attention mechanism is essential for the complex tasks which require capturing multiple different semantic relationships between words.

3.3 Large Language Models

Large language models (LLMs) are a recent advancement built on top of the transformer architecture for natural language processing, demonstrating state of the art capabilities in understanding, reasoning and text generation. These models form the core of most modern NL2SQL systems like XiYan-SQL, DAIL and others. They have a wide language understanding as they have been trained on large text corpora, are exposed to different domains of traning data and have an inherent understanding of the SQL syntax as they have been exposed to it during pre-training. This shifted the approach while developing NL2SQL systems to focus on in-context learning methods instead of training dedicated models from scratch.

3.3.1 Feedforward and Residual Connections

Every layer in a transformer combines a multi-head attention with position-wise feedforward networks:

$$\text{FFN}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2 \tag{6}$$

Residual connections and layer normalization stabilize training of complex networks:

$$\text{Output} = x + \text{Sublayer}(\text{LayerNorm}(x)) \tag{7}$$

where Sublayer represents either multi-head attention or the feedforward network. These components enable transformers to scale to billions of parameters while maintaining training stability.

3.3.2 Positional Encoding

Since transformers are able to process sequences in parallel, positional encodings inject information about token positions. Original transformers use sinusoidal functions, while modern LLMs often employ learned positional embeddings or relative position encodings. This positional information is inherently important for NL2SQL systems, where the order of generated SQL clauses (e.g., `SELECT` must come before before `WHERE`) carries significant semantic meaning.

3.3.3 Pre-training and Fine-tuning

Large language models are trained in a two-fold process that separates the training of general language understanding from from task-specific capabilities.

3.3.3.1 Pre-training

During pre-training LLMs are exposed to vasts amounts of unlabeled text gathered from the internet, books etc. Using self-supervised objectives, no manual labeling of the datasets is required. The most common pre-training objective is causal language modeling, where a model learns to predict the next token given a previous context:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t}; \theta) \tag{8}$$

This objective ingrains general language patterns, common knowledge, reasoning capabilities, and even programming knowledge into models when code is included in the training datasets. Models like GPT, LLaMA, and Qwen are pre-trained using this objective, aiming to predict coherent continuations of text sequences.

3.3.3.2 Fine-tuning

Fine-tuning is the process of adapting a pre-trained model, for example for solving NL2SQL tasks, by continuing model training on a labeled dataset (eg, natural language question to expected sql query). During fine-tuning, the model can learn to map natural language questions to expected SQL queries when being presented with a question and a database schema.

Fine-tuning usualy require significantly smaller datasets (thousands to hundreds of thousands of examples) compared to pre-training (billions to trillions of tokens), largely relying on the natural language understanding capabilities acquired during the pre-training phase.

Models like OmniSQL demonstrate that smaller fine-tuned models can outperform significantly larger LLMs for NL2SQL tasks using a fraction of the parameters (H. Li et al., 2025).

3.3.4 In-Context Learning

In-context learning (ICL) is the process of teaching a model how to perform a certain task by providing examples in its prompt, without updating models parameters. Thus ICL and Fine-tuning differ by their timing: runtime vs learning.

This capability emerged with sufficiently large models which have large enough context windows to fit in examples of solutions for similar problems. This method was shown to be very effective by **?** in **?**, improving the performance of general purpose LLMs as well as fine-tuned LLMs when being presented with relevant examples (**?**). The combination of fine-tuned models like OmniSQL with ICL is therefore a promising research for NL2SQL systems.

3.3.5 Prompt Engineering

Prompt engineering is the process of structuring the input to LLMs in a way where they adhere to desired behaviours. For NL2SQL systems previous research has found diverging effectiveness between different example, schema and question presentation mechanisms (H. Li et al., 2025; **?**).

3.3.6 Chain-of-Thought Reasoning

Prompting or training models to output their chain-of-thought encourages models to generate reasoning traces before producing their final answer to a question or input. For NL2SQL, this usually involves generating explanations of query logic, analysing the input database schema, and planning the query structure before generating a query.

Fine-tuned models like OmniSQL are fine-tuned to produce chain-of-thought output during generation in markdown which enables debugging model behaviour and understanding the query generation approach.

3.3.7 Model Limitations

Understanding the limitations of large language models is essential for designing robust NL2SQL systems.

3.3.7.1 Hallucination

As LLMs are just predicting a sequence of output tokens, they have no inherent understanding of the actual problem domain and solution space they are presented with. Thus in NL2SQL systems, LLMs might generate invalid SQL queries which might reference non existent tables, columns or relationships, output invalid syntax or generate promising but semantically invalid queries.

These flaws are inherent to a text based representation and are non recoverable, although through candidate validation and self-refinement approaches most obviously invalid queries can be recovered. Detecting semantic invalidity is a hard research problem on its own.

3.3.7.2 Context Window Limitations

As the underlying transformer architecture is still bound to finite context windows, extensive database schemas or long lists of examples can exceed the window, thus causing the query generation to fail.

3.4 Natural Language Processing

Natural language processing (NLP) provides the foundations for working with natural language inputs in computer science and mathematics. It allows for text understanding and text representation, aswell as text processing and semantic feature extraction, capturing meaning and structure. For NL2SQL systems NLP enable semantic understanding of natural language queries and semantic search using linguistic similarity. Using NLP and text embeddings reference datasets can be used for semantic search effectively and accurately.

3.4.1 Text Preprocessing and Tokenization

Before neural networks can process text, raw text must be converted into a structured representations that models can process. This transformation is called tokenization, which refers to the process of segmenting text into discrete units (tokens) for subsequent processing.

3.4.1.1 Tokenization Approaches

Tokens are short character chains that allow representing text efficiently as a series of tokens by refering to the tokenizers vocabulary. Therefore NLP systems mostly employ subword tokenization. Modern tokenization approaches balance vocabulary size with representational efficiency. Common approaches include **Byte-Pair Encoding (BPE)**, **WordPiece** and **SentencePiece**.

These subword tokenization approaches are particularly valuable for NL2SQL because database schemas often contain domain-specific terminology, compound words, and technical terminology that may not appear in common vocabularies. Subword approaches handle unseen words gracefully by decomposing them into meaningful tokens.

3.4.1.2 Tokenization in NL2SQL Contexts

For NL2SQL systems, tokenization must be capable of both encoding natural language queries and SQL queries. SQL presents a challenge in NLP as it relies on a structured syntax and usage of special characters (dots, underscores, SQL operators, etc.) which are less commonly found in natural language.

3.4.2 Word Embeddings and Semantic Representations

Word embeddings are the vector space representation of a tokens where semantic meaning is captured through geometric properties. As word embeddings are ordinary vectors in a high-dimensional vector space, common vector operations can still be performed.

3.4.2.1 Embedding Functions

An embedding function maps tokens from a discrete vocabulary V to dense vectors in $\mathbb{R}^d$:

$$e_w : V \to \mathbb{R}^d \tag{9}$$

where d is the embedding dimensionality (usually between 256-8192 for modern models). The key property of the vector space $\mathbb{R}^d$ is that semantically similar words occupy nearby regions in this vector space, hence vector operations and proximity in the vector space allows for computation of semantic properties such as semantic distance.

Early embedding methods like Word2Vec and GloVe learned static representations where each word has a single vector regardless of its surrounding context. Modern contextual embeddings, generate different vectors for the same word based on its context, thus capturing meaning more accurately.

3.4.2.2 Sentence Embeddings

While word embeddings are useful for determining the relationships between words, they are limited in length and complexity. NL2SQL systems require representations of entire questions and SQL queries in order to discover answers to similar questions asked in the past. Sentence embeddings aggregate the token-level meaning into fixed-size vectors in the same vector space $\mathbb{R}^d$, thus representing complete sentences or texts in one vector.

$$e_s : (v_0, .., v_n) \to \mathbb{R}^d \tag{10}$$

Different aggregation / pooling strategies emerged, like mean pooling, max pooling, using a CLS token or learned aggregation. Natural uses an embedding model (Qwen3-Embedding-8B) trained specifically for generating high-quality sentence representations for subsequent searching.

3.4.3 Semantic Similarity Metrics

Quantifying the similarity between two natural language sentences is essential for example selection in NL2SQL. This requires that similarity metrics in the vector space align with natural understanding of semantic relation.

3.4.3.1 Cosine Similarity

Cosine similarity is a measurement using the angle between two vectors, to capture their directional similarity while normalizing for magnitude:

$$s_{\cos} = \frac{u \cdot v}{\|u\|\|v\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \tag{11}$$

Thus cosine similarity ranges from -1 which indicates opposite directions to 1 representing same direction, with 0 indicating orthogonality of the input vectors. This metric is widely used in NLP as it is independent of vector magnitude, efficient to compute and well-suited for high-dimensional vector spaces.

3.4.3.2 Cosine Distance

Cosine distance converts similarity into a distance metric:

$$d_{\cos}(u, v) = 1 - \mathrm{s}_{\cos}(u, v) \tag{12}$$

Thus cosine distance ranges from 0 for identical direction to 2 for opposite direction input vectors. This is satisfying properties desirable for distance metrics while preserving the angular relationship of the input vectors u and v.

3.5 Graph Theory

Graph theory is a mathematical domain for working with structures and relationships. It is highly relevant for databases and NL2SQL systems as modeling relationships in databases maps intuitively to graphs. This section introduces the theoretical framework of graphs, graph kernels and graph similarity which are required foundations for the example selection algorithm of NATURAL.

3.5.1 Graph Fundamentals

A graph is noted as $G = (V, E)$ in its simplest form and consists of vertices V and edges $E \subseteq V \times V$ which are connections between vertices. Labeled graphs are an extension to this concept which associate labels with vertecies and edges, encoding semantic information such as vertex types, attributes, or constraints.

3.5.2 Graph Kernels

Graph kernels are functions $k : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ which are measuring the structural similarity between two graphs:

$$k(G_1, G_2) = \langle \phi(G_1), \phi(G_2) \rangle \tag{13}$$

where $\phi : \mathcal{G} \to \mathcal{H}$ maps graphs to a feature space $\mathcal{H}$, which allows measuring similarity in the feature space, rather than using raw graph matching to compute the similarity. Thus, on a high level, graph kernels may be used to identify common super structures in graphs.

3.5.3 Weisfeiler-Lehman Algorithm

The Weisfeiler-Lehman (WL) algorithm is a graph kernel algorithm which iteratively refines node labels based on the neighborhood structure in the graph (Togninalli, Ghisu, Llinares-López, Rieck, & Borgwardt, 2019):

1. **Sorting** — Represent vertices as a sorted list of neighbors (L_v)
2. **Compression** — Compute the hash of L_v for every vertex ($h(L_v)$)
3. **Relabeling** — Relabel every vertex v with $h(L_v)$ as its new node label

This algorithm is repeated for w iterations (typically $w = 3 - 5$, due to deminishing returns). After the iterative vertex label computation, labels encode the w-hop neighborhood structure. Subsequently a label histogram can be used as the graph-level feature vector, enabling efficient structural comparison.

3.5.4 Wasserstein Distance

The Wasserstein distance measures the minimum cost of transforming one probability distribution into another (also known as earth mover's distance). Given two distributions μ and ν over space $\mathcal{X}$ with metric d:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, d\gamma(x, y) \right)^{1/p} \tag{14}$$

where $\Gamma(\mu, \nu)$ is the set of couplings of μ and ν. Unlike other distance metrics of probability distributions like the Kullback-Leibler divergence, the Wasserstein distance accounts for the underlying metric space structure (Togninalli et al., 2019).

3.5.5 Wasserstein Weisfeiler-Leman Kernels

WWL kernels combine the labels computed using WL with Wasserstein distance to measure graph distance (Togninalli et al., 2019):

1. Apply w WL iterations to both graphs, producing node label distributions at each iteration
2. Compute Wasserstein distance between graphs using Hamming distance (for the categorical features) or Euclidean distance (for continuous features) as the ground metric
3. Aggregate distances into a similarity score using a Laplacian kernel:
   $k_{\text{WWL}}(G_1, G_2) = \exp(-\lambda D_W))$

The resulting graph kernel captures both local neighborhood structures (using the WL algorithm) and global distributional properties (using the Wasserstein distance) (Togninalli et al., 2019). For database schemas, WWL kernels can similar structural patterns (table structures, possible joins, constraints) while remaining robust to minor structure variations. This makes them particularly well suited for NL2SQL example selection algorithms that factor in schema similarity.

4 System Design

This chapter describes the design of NATURAL, our proposed NL2SQL system, addressing limitations and research gaps identified in the literature review.

NATURAL is architected as a pipeline that transforms natural language questions into SQL queries using example selection (σ), schema subsetting (ϕ), and self refinement (ρ) and voting (ν). The proposed system consists of five components:

1. **Example Selection** σ – Identifies semantically and structurally similar examples using cosine similarity and schema distance.
2. **Schema Subsetting** ϕ – Reducing schema complexity by subsetting the schema to the relevant subset of tables and relationships.
3. **Query Projection** π – Few-shot learning with a finetuned model to project natural language queries to SQL statements.
4. **Self Refinement** ρ – Self refinement of generated SQL statements through execution feedback and error analysis.
5. **Voting** ν – Self consensus voting mechanism to choose the most likely result from multiple generation attempts.

The system processes queries using the following algorithm:

```
  Natural Language Query
→ Sketch Generation
→ Example Selection
→ k-times [
    → Schema Subsetting
    → Few-Shot Generation
    → Self-Correction
  ]
→ Consensus Voting
→ SQL Query
```

This design largely builds upon few-shot learning concepts from DAIL-SQL (D. Gao et al., 2023) and OmniSQL (H. Li et al., 2025) but incorporates novel contributions in schema-aware example selection by harnessing a graph-based structural similarity metric.

4.1 Initialization

Before the NATURAL system can process queries, it requires initialization with historical data to construct a specific embedding space $v \in \mathcal{V}$ and schema distance index $d \in \mathcal{D}$. This preprocessing phase analyzes existing datasets and previous user interactions to enable semantically-aware example selection.

The initialization process consists of two main components: (1) embedding generation for semantic similarity computation, and (2) schema indexing for structural similarity measurement. These components work together to support the example selection function σ.

4.1.1 Embedding

NATURAL constructs an embedding space v to enable semantic similarity search (D. Gao et al., 2023). The system uses cosine similarity to identify relevant examples based on both the natural language queries and SQL statements, allowing efficient retrieval from large collections of question-answer pairs.

We therefore embed a set of training samples $\mathcal{T} = \{(q_1, \omega_1), ..., (q_n, \omega_n)\}$ into an embedding space v, where each q_i is a natural language question and ω_i is the corresponding SQL query.

The embedding space v can be formally defined as:

$$v = \{\ (q, \iota(q), \omega, \iota(\omega)) \mid (q, \omega) \in \mathcal{T}\ \}$$

where ι is the embedding function that maps text to vector representations.

4.1.2 Schema Indexing

To choose the most relevant subset of samples for few-shot learning, it is important that the SQL queries we choose as examples are written for structurally similar database schemas in order to minimize the structural difference between the selected samples and the ground truth query for a given natural language question.

For example ω_{ground} the question "Give me all contacts for the user with the id 10" might look different depending on the database schema at hand, thus only selecting samples based on the similarity of the natural language question will yield inferior sample quality.

```
CREATE TABLE users (
  id TEXT PRIMARY KEY
);

CREATE TABLE contacts (
  user_id TEXT NOT NULL,
  name TEXT NOT NULL,
  FOREIGN KEY (user_id)
   REFERENCES users(id),
  PRIMARY KEY (user_id, name)
);
```

Figure 1: Normalized schema

```
CREATE TABLE users (
  id TEXT PRIMARY KEY,
  contacts TEXT[] NOT NULL
);
```

Figure 2: Denormalized schema

Given the database schemas in figures 1 and 2 respective definitions of ω_{ground} would be 3 and 4.

```
SELECT contacts.name
FROM users
JOIN contacts
  ON contacts.user_id = users.id
WHERE users.id = 10;
```

Figure 3: SQL JOIN selection

```
SELECT contacts
FROM users
WHERE id = 10;
```

Figure 4: SQL Array selection

As shown in the figures 3 and 4 the structural similarity of the underlying database schema is a crucial component of the relevance of an example.

4.1.2.1 Graph Representation

To determine structural similarity of database schemas systematically, we propose using graph representation of database schemas as a data definition language (DDL) independent representation of the database structure. Choosing a graph representation allows us to leverage established methods for measuring graph similarity such as Wasserstein-Weisfeiler-Lehman graph kernels (Togninalli et al., 2019). Given a database schema s we define the corresponding graph $G_s = (V, E, \ell, w)$ as:

1. $V = V_t \cup V_c$ where V_t represents table nodes and V_c represents column nodes
2. $E = E_{tc} \cup E_{fk} \cup E_{ref}$ where:
   - E_{tc} are table-column edges
   - E_{fk} are foreign key relationship edges between tables
   - E_{ref} are reference edges between foreign key columns

3. Each node $v \in V$ has a semantic label $\ell(v) \in \{1, 2, \ldots, 9\}$ where:
   - $\ell(v) = 1$ for table nodes
   - $\ell(v) = 2$ for generic column nodes
   - $\ell(v) = 3$ for primary key columns
   - $\ell(v) = 4$ for foreign key columns
   - $\ell(v) = 5$ for text columns
   - $\ell(v) = 6$ for numeric columns
   - $\ell(v) = 7$ for datetime columns
   - $\ell(v) = 8$ for boolean columns
   - $\ell(v) = 9$ for other/unknown data types
4. Each edge $e \in E$ has a weight $w(e) \in \{0.5, 0.9, 1.0\}$ reflecting its structural importance:
   - $w(e) = 1.0$ for foreign key relationships (highest importance)
   - $w(e) = 0.9$ for column foreign key reference edges
   - $w(e) = 0.5$ for table-column edges

We define the structural similarity of two databases as the topological distance of the respective graphs. The graph representation captures essential schema structure through semantic node labeling that prioritizes constraints over data types: primary key columns receive label 3 regardless of their underlying data type, foreign key columns receive label 4, and only then are remaining columns categorized by data type (text=5, numeric=6, datetime=7, boolean=8, other=9). This hierarchical labeling ensures that structural relationships take precedence over exact content, while omitting table names, column names and domain terminology to achieve schema-agnostic comparison.

The graph representation of the database schemas introduced in 1 and 2 are therefore:

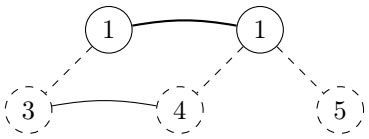

Figure 5: Normalized graph repr.

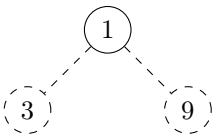

Figure 6: Denormalized graph repr.

Legend: ○ table ◌ column **-** foreign key – reference - - table-column

The δ of the order of the graphs 5 and 6 further highlight the structural difference between the two schemas.

4.1.2.2 Distance Measurement

To measure the distance between the graphs displayed in 5 and 6, we employ the Wasserstein Weisfeiler-Lehman (WWL) graph kernel method (Togninalli et al., 2019).

WWL kernels compute structural similarity by combining discrete Weisfeiler-Lehman graph features with continuous optimal transport theory. The method first extracts WL features from the labeled graph structure, then computes the Wasserstein distance between the resulting feature distributions. For two graphs G and G' with semantic node labels, the distance is computed as:

$$D_W^f(G, G') = W_1(f(G), f(G')) \tag{15}$$

as introduced by Togninalli et al. in 2019.

For the database schemas above, the graph representation of the normalized schema G_{norm} (see figure 5) with 5 nodes and 5 edges and denormalized schema G_{denorm} (see figure 6) with 3 nodes and 2 edges show significant topological differences. The WWL distance captures both the reduction in graph complexity (fewer nodes and edges) and the loss of relational structure (elimination of foreign key relationships), resulting in a substantial distance value reflecting their structural dissimilarity despite representing equivalent logical data with $D_W^f(G_{norm}, G_{denorm}) \approx 0.78$.

4.1.2.3 Distance Index

The distance index d maintains precomputed distances between all observed database schemas, enabling efficient retrieval of structurally similar databases during sample selection. This index is constructed as a set of schema-distance tuples $\{(sm_1, di_1), \ldots, (sm_i, di_i)\}$ where sm_i is a database schema name and di_i denotes its WWL distance to the current schema.

4.2 Functions

The following sets are subsequently used to introduce the functions σ, ϕ, π, ρ and ν.

1. $\mathcal{Q}$ – The set of all possible natural language queries.
2. $\mathcal{S}$ – The set of all possible database schemas.
3. $\mathcal{E}$ – The set of all possible execution functions for SQL validation.
4. $\mathcal{V}$ – The set of all possible embedding spaces.
5. $\mathcal{D}$ – The set of all possible distance indices over historically observed databases.
6. $\mathcal{C}$ – The set of all possible candidate sets $\mathcal{C}'$.
7. $\mathcal{Q}_{\mathcal{S}}$ – The set of all valid queries over the database schema S.

4.2.1 Example Selection – σ

The selection function σ retrieves the most relevant examples from the historical embedding space v based on semantic similarity to the input query. This function implements the core example selection mechanism that subsequently enables few-shot learning for SQL generation.

$$\sigma : \mathcal{Q} \times \Omega_{\mathcal{S}} \times \mathcal{S} \times \mathcal{V} \times \mathcal{D} \to \mathcal{C}'$$

For a specific input instance, we write:

$$\sigma(q, z, s, v, d) = \{(q_1, s_1, \omega_1, d_1), \ldots, (q_k, s_k, \omega_k, d_k)\}$$

where $q \in \mathcal{Q}$ is the input natural language query, $z \in \Omega_{\mathcal{S}}$ is the zero-shot inference result, $s \in \mathcal{S}$ is the database schema, $v \in \mathcal{V}$ is the embedding space, $d \in \mathcal{D}$ is the distance index, and the result is a candidate set $\mathcal{C}' \in \mathcal{C}$ containing k tuples of natural language queries, schemas, corresponding SQL queries and their combined distance.

The function utilizes cosine similarity in the embedding space to identify examples that are semantically closest to the input query q, considering both natural language representations and schema distance as described in Section 4.1.1. This approach ensures that examples are weighted by the following properties:

1. **Question similarity** – Semantic similarity between the input query and historically observed natural language queries in the embedding space v, measured using cosine distance between their respective embeddings.
2. **SQL similarity** – Structural similarity between the zero-shot inference result z and candidate SQL queries using code representations, enabling pattern recognition across different database domains.
3. **Database similarity** – Schema compatibility measured through the distance index d, ensuring selected examples operate on analogous database structures with similar table relationships and column types.

The exact weighting of these can be adjusted through three constants, w_q, w_s, and w_d.

Algorithm 1 σ - Example Selection

Require: $q \in \mathcal{Q}$, $z \in \Omega_{\mathcal{S}}$, $s \in \mathcal{S}$, $v \in \mathcal{V}$, $d \in \mathcal{D}$
Require: $k \in \mathbb{N}$, $k \geq 1$ ▷ Number of examples to select
$candidates \leftarrow \emptyset$ ▷ Initialize candidate set
for $(q_i, s_i, \omega_i) \in v$ **do** ▷ For each sample
  $sim_q \leftarrow cosine(\iota(q), \iota(q_i))$ ▷ Calculate question similarity
  $sim_s \leftarrow cosine(\iota(z), \iota(\omega_i))$ ▷ Calculate SQL similarity
  $sim_d \leftarrow d(s, s_i)$ ▷ Calculate schema distance
  $score \leftarrow w_q \cdot sim_q + w_s \cdot sim_s + w_d \cdot sim_d$
  $candidates \leftarrow candidates \cup \{(q_i, s_i, \omega_i, score)\}$
end for
return $top_k(candidates)$ ▷ Return k highest scoring examples

4.2.2 Schema Subsetting – ϕ

The subsetting function ϕ reduces the database schema to only include tables, columns, and relationships that are relevant to the current set of candidates C. This schema pruning mechanism reduces the complexity of the SQL generation task and token efficiency by focusing on the most relevant schema elements.

$$\phi : \mathcal{C}' \times \mathcal{S} \to \mathcal{S}'$$

For a specific input instance, we write:

$$\phi(c, s) = s'$$

where $c \in \mathcal{C}'$ is a candidate set containing selected examples, $s \in \mathcal{S}$ is the full database schema, and $s' \subseteq s$ is the reduced schema subset containing only relevant tables and columns.

The function analyzes the selected examples in c to identify which schema elements are commonly referenced in similar queries. This analysis enables the system to steer the model's attention on the most relevant portions of potentially large and complex database schemas, thereby improving both accuracy and computational efficiency and prevents exceeding limited context windows.

Algorithm 2 ϕ - Schema Subsetting

Require: $c \in \mathcal{C}'$, $s \in \mathcal{S}$
$s' \leftarrow \emptyset$ ▷ Empty schema subset
for $table \in s$ **do**
  if $\exists\ r \in references(c, table)$ **then** ▷ Any candidates reference the table
    $s' \leftarrow s' \cup \{table\}$ ▷ Extend subsetted schema
  end if
end for
return s' ▷ Return subsetted schema

4.2.3 Query Projection – π

The projection function π represents the core translation mechanism that converts natural language queries into SQL statements using the selected examples and database schema. This function encapsulates the inference using LLMs that generates candidate SQL queries based on the provided context.

$$\pi : \mathcal{Q} \times \mathcal{S} \times \mathcal{C}' \to \Omega_{\mathcal{S}}$$

For a specific input instance, we write:

$$\pi(q, s, c) = \omega$$

where $q \in \mathcal{Q}$ is the input natural language query, $s \in \mathcal{S}$ is the database schema, $c \in \mathcal{C}'$ is the set of selected examples, and $\omega \in \Omega_{\mathcal{S}}$ is the generated SQL query.

The function operates by constructing a prompt that combines the natural language query, the relevant schema information, and the selected examples in a format optimized for large language model inference. The model then generates a SQL query that attempts to capture the semantic intent of the natural language input while adhering to the constraints imposed by the database schema.

Algorithm 3 π - Query Projection

Require: $q \in \mathcal{Q}$, $s \in \mathcal{S}$, $c \in \mathcal{C}'$
Require: $\tau \in \mathbb{R}$, $\tau > 0$ ▷ Relevance threshold
$c_{rel} \leftarrow \{ (q_i, s_i, \omega_i, d_i) \in c : d_i < \tau \}$ ▷ Filter by relevance
$ex \leftarrow \{ fmt(q_i, \omega_i, d_i) \mid (q_i, s_i, \omega_i, d_i) \in c_{rel} \}$ ▷ Format examples
$prompt \leftarrow prompt(q, s, ex)$ ▷ Construct prompt
$out \leftarrow model(prompt)$ ▷ Model inference
$\Omega_{cand} \leftarrow extract_{sql}(out)$ ▷ Extract SQL candidates
for $\omega_{raw} \in \Omega_{cand}$ **do** ▷ Validate candidates
  if $\omega_{raw} \in \Omega_{\mathcal{S}}$ **then** ▷ Check syntactic validity
    return ω_{raw} ▷ Return first valid query
  end if
end for

4.2.4 Self Refinement – ρ

The refinement function ρ implements a self-correction mechanism that iteratively improves generated SQL queries by identifying and correcting syntax errors, semantic inconsistencies, and execution failures. This function enables the system to learn from its mistakes and produce higher-quality outputs through automated feedback loops.

$$\rho : \mathcal{Q} \times \mathcal{S} \times \mathcal{E} \times \Omega_{\mathcal{S}} \to \Omega_{\mathcal{S}}$$

For a specific input instance, we write:

$$\rho(q, s, e, \omega_{raw}) = \omega_{refined}$$

where $q \in \mathcal{Q}$ is the original natural language query, $s \in \mathcal{S}$ is the database schema, $e \in \mathcal{E}$ is the execution function for validation, $\omega_{raw} \in \Omega_{\mathcal{S}}$ is the previously generated SQL query and $\omega_{refined} \in \Omega_{\mathcal{S}}$ is the improved SQL query.

The function operates by executing the candidate query against the database, analyzing any errors or unexpected results, and then prompting the language model to generate an improved version based on the identified issues. This iterative refinement process continues until a valid, executable query is produced or a maximum number of refinement attempts is reached.

Algorithm 4 ρ - Self Refinement

Require: $q \in \mathcal{Q}$, $s \in \mathcal{S}$, $e \in \mathcal{E}$, $\omega_{raw} \in \Omega_{\mathcal{S}}$
$exec \leftarrow e(s, \omega_{raw})$ ▷ Verify execution output
$prompt \leftarrow prompt_{refine}(q, s, \omega_{raw}, exec)$ ▷ Construct refinement prompt
$out \leftarrow model(prompt)$ ▷ Generate refinement output
$\Omega_{ref} \leftarrow extract_{sql}(out)$ ▷ Extract refined candidates
for $\omega_{ref} \in \Omega_{ref}$ **do** ▷ Validate refined queries
  if $\omega_{ref} \in \Omega_{\mathcal{S}}$ **then** ▷ Check syntactic validity
    return ω_{ref} ▷ Return first valid refinement
  end if
end for

4.2.5 Voting – ν

The voting function ν implements a consensus mechanism that selects the most reliable SQL query from multiple candidate solutions generated through the pipeline. This function enhances robustness by leveraging the result distribution of multiple generation attempts to identify the most likely correct answer.

$$\nu : \mathcal{C} \times \mathcal{E} \to \Omega_{\mathcal{S}}$$

For a specific input instance, we write:

$$\nu(C, e) = \omega_{consensus}$$

where $C \in \mathcal{C}$ is a set of candidate SQL queries, $e \in \mathcal{E}$ is the execution function for validation, and $\omega_{consensus} \in \Omega_{\mathcal{S}}$ is the selected consensus query.

The voting function ν implements a majority voting algorithm similar to that described by OmniSQL (H. Li et al., 2025), where the result that appears most frequently across multiple generation attempts is deemed to be the most likely correct answer. The function applies frequency-based selection among the valid candidates. In cases where no clear majority exists, the function may apply additional heuristics such as query complexity or execution performance to determine the final proposed query candidate $\omega_{consensus}$.

Algorithm 5 ν - Consensus Voting

Require: $C \in \mathcal{C}$, $e \in \mathcal{E}$

$results \leftarrow \{\}$ ▷ Map from result sets to candidate lists
for $\omega \in C$ **do** ▷ Execute each candidate
  $result \leftarrow e(\omega)$ ▷ Execute query
  $results[result] \leftarrow results[result] \cup \{\omega\}$
end for
$dist \leftarrow \{results[r] : r \in results\}$ ▷ Get all result groups
$dist \leftarrow sort(dist, |group| \mapsto |group|)$ ▷ Sort by group size (largest first)
return $top_1(top_1(dist))$ ▷ Return largest group

4.3 Composition – nq

The nq function composes the system by first establishing a zero-shot baseline, then using the precomputed embedding space v and distance index d from the initialization phase to guide few-shot generation while self refining results through execution feedback and finally yield the most likely candidate ω through majority voting.

$$nq : \mathcal{Q} \times \mathcal{S} \times \mathcal{E} \times \mathcal{V} \times \mathcal{D} \to \Omega_{\mathcal{S}}$$

For a specific input instance, we write:

$$nq(q, s, e, v, d) = \omega$$

where $q \in \mathcal{Q}$ is the input natural language query, $s \in \mathcal{S}$ is the database schema, $e \in \mathcal{E}$ is the execution function for validation, $v \in \mathcal{V}$ is the embedding space, $d \in \mathcal{D}$ is the distance index, and $\omega \in \Omega_{\mathcal{S}}$ is the generated SQL query.

The *nq* function implements Algorithm 6.

Algorithm 6 *nq*

Require: $q \in \mathcal{Q}$, $s \in \mathcal{S}$, $e \in \mathcal{E}$, $v \in \mathcal{V}$, $d \in \mathcal{D}$
Require: $k \in \mathbb{N}$, $k \geq 1$ ▷ Number of candidates to generate
$z \leftarrow \pi(\emptyset, q, s)$ ▷ Generate zero-shot baseline
$\mathcal{C}' \leftarrow \sigma(q, z, s, v, d)$ ▷ Select relevant examples
$C \leftarrow \emptyset$ ▷ Initialize candidate set
while $|C| < k$ **do** ▷ Generate k candidates
    $s' \leftarrow \phi(\mathcal{C}', s)$ ▷ Subset schema to relevant elements
    $\omega \leftarrow \pi(\mathcal{C}', q, s')$ ▷ Generate candidate query
    if $e(\omega, s')$ **then** ▷ Validate syntactic correctness
        $C \leftarrow C \cup \{\omega\}$ ▷ Add valid candidate
    end if
    $\omega' \leftarrow \rho(q, s', e, \omega)$ ▷ Attempt self-correction
    if $e(\omega', s')$ **then** ▷ Validate corrected query
        $C \leftarrow C \cup \{\omega'\}$ ▷ Add corrected candidate
    end if
end while
return $\nu(C, e)$ ▷ Select consensus result

5 Implementation

This chapter presents the practical implementation of the NATURAL system, translating the theoretical framework outlined in Chapter 4 into a working NL2SQL system.

Primary attention is given to the engineering challenges encountered, performance bottlenecks identified, and optimization strategies implemented to achieve practical deployment viability on consumer hardware. The chapter is comprised of a software architecture and infrastructure discussion as well as implementation outlines of each pipeline component.

5.1 Architecture and Infrastructure

The implementation of the system design phase is split into inference, sampling and evaluation code. The runtime code focuses on the actual algorithm implementations outlined in section 4, sampling code focuses on indexing samples and computing distance indices and the evaluation code runs the NATURAL pipeline on prevalent benchmarks.

This subsection focuses on outlining the scope of each technological component, the design rationale behind them and discusses the technology stack decisions.

5.1.1 Software Architecture

NATURAL consists of five software components:

1. `natural-models` – For handling the actual model execution on GPUs for inference and embedding.
2. `natural-graphs` – Graph library for representing database schemas in graphs as well as computing graph similarities and distances.
3. `natural-inference` – The core pipeline implementation using other software components during inference time.
4. `natural-sampling` – The sampling setup that focuses on indexing samples and computing graph distance indices.
5. `natural-benchmark` – The benchmarking setup used to run experiments, continuously evaluate the accuracy of the system and compile statistics.

```
THIS NEEDS A DIAGRAM TO UNDERSTAND THE SOFTWARE ARCH.
```

This separation of concerns emerged as the split between inference, sampling and evaluation code required shared fundamentals like model execution and graph representation which in turn improves testability and maintainability. Furthermore separating the inference, sampling and evaluation code yields a smaller footprint when embedding the inference code into databases.

Challenges associated with this multi component architecture are mostly tied to dependency management and increased build complexity although these can be mitigated through `cargo`'s workspace support.

5.1.2 Resource Management Strategy

Given the constrained 24 gigabytes of VRAM capacity of the RTX 3090 not all available model sizes can be used. While 14B model variants theoretically work, the need for an embedding model typically exhausts the VRAM unless using steep quantization formats. The OMNISQL 7B models were shown to have an average performance degradation of only 0.3% compared to their larger 14B counterparts by H. Li et al. in 2025. On the spider test dataset the 7B model surprisingly outperformed its 14B counterpart by 0.6%. Due to the unclear performance gains of the 14B model, the significant increase in inference time and the limited VRAM available, this thesis focuses on the 7B variant of OMNISQL.

As an embedding model is required for doing semantic search of samples in the pipeline a small companion model is required to embed both the sample datasets (SYNSQL, BIRD and SPIDER) during sampling time and the user provided natural language question during inference. The Qwen3 series embedding models ranked first place on the MTEB multilingual leaderboard. With the Q8_0 quantized version of the 8B model consumes 8 gigabytes of VRAM which is not possible to fit into 24 gigabytes of VRAM when accounting for KV-Cache

requirements. Therefore, instead of choosing a heavily quantized version of the 8B variant (ie, Q4_KM or below), the 4B model with Q8_0 shows similar performance characteristics on simple use cases while offering a significantly smaller memory footprint at 2 gigabytes. https://huggingface.co/Qwen/Qwen3-Embedding-8B cite.

As models have significant loading times, the system loads models globally and at startup and hands around references using atomic reference counting (`std::sync::Arc`) to ensure that models are not loaded twice, can be reused between different inference calls and GPU memory is not exhausted which would lead to a program crash.

5.1.3 Technology Stack Decisions

The choice of programming language, inference frameworks, and supporting libraries implies the system's performance and deployment characteristics, as well as development velocity. This section analyzes the key decisions made for NATURAL, discussing their trade-offs between research flexibility and production readiness, performance optimization and development speed, and ecosystem maturity versus cutting-edge capabilities.

The critical decisions are the programming language, the model inference framework, and the vector similarity search solution. Each decision was evaluated against the constraints of limited hardware resources (24GB VRAM), the need for database integration capabilities, and the requirement for reproducible research outcomes.

5.1.3.1 Language and Ecosystem

The most apparent and impactful decision is likely the language and ecosystem choice made. Viable languages for implementing natural language processing and machine learning heavy systems are Python, R, Julia, Rust, C / C++ and Java as well as other general purpose languages.

While interpreted languages like Python, R and Julia tend to be significantly easier to use for rapid prototyping and approach validation due to their loose type systems and great scientific ecosystem, they come with serious drawbacks with regards to deployability, speed and robustness compared to compiled and strongly typed languages.

Languages like Java, C / C++ and Rust offer greater stability at runtime, better interoperability into other programs (eg, database extensions) and better resource utilization they represent an interesting trade-off between performance optimization and portability vs. speed of iteration and research ecosystems.

R and Julia are inferior to Python when it comes to adoption, machine learning frameworks and natural language processing. Java requires a Java Virtual Machine (JVM) at runtime which yields worse portability than C / C++ and Rust while offering little advantage over interpreted languages like Python. Thus Python, C / C++ and Rust emerge as strong contenders for the implementation of NATURAL. C and C++ have the primary downside that they are prone to program crashes and memory safety issues whereas Rust resolves most of these downsides while maintaining similar performance, memory management and portability characteristics.

Given that performance plays a critical role when developing machine learning systems with limited access to hardware, the path to a potential production deployments of NL2SQL is easier and the language ecosystem offers bindings for the most notable scientific libraries, Rust offers a great value proposition for ML systems at the cost of development speed.

5.1.3.2 Inference Framework

For local model inference, two primary frameworks emerged: llama.cpp (FFI), Candle (Rust-native). While Rust-native frameworks generally offer better portability, llama.cpp provided a better balance between performance, ecosystem maturity, and implementation simplicity.

llama.cpp provides out-of-the-box optimizations with comprehensive GGUF quantization support, yielding 25-30 tokens per second on consumer grade RTX 3090 hardware. It offers high-level abstractions for model loading and tokenization and sampling. The mature GGML ecosystem and advanced quantization schemes (like `Q4_K_M`, `Q8_0`) justified the FFI complexity.

Candle, even though it offered comparable raw inference speeds, required rather extensive low-level implementation work and proved incompatibility with scenarios where the resulting binary is embedded into a database (e.g. PostgreSQL). The combination of manual tensor management, more convoluted quantization support, and deployment constraints made llama.cpp a better alternative for production-ready systems.

5.1.3.3 Similarity Search Framework

For lightweight similarity search SQLite in combination with it's `sqlite-vec` extension is a sensible option. Especially for smaller data loads introducing the complexity of `faiss` and comparable vector search solutions like `qdrant` outweighs their speed benefits.

The `sqlite-vec` extension advertises to be a "fast enough" vector search solution, allowing for reduced complexity and compatibility with graphical database interfaces that support SQLite to inspect the embedding space.

5.1.3.4 Trade-offs

Overall the development overhead of choosing Rust over Python for the implementation phase was noticeable; Rust's machine learning and natural language processing frameworks are less advanced, porting research code written in python from other papers (like H. Li et al. (2025)) turned out to be non trivial. This decision likely doubled the time needed for the implementation, but in turn provides a clear path for the algorithms in this thesis to be productionized. The outcome of the implementation has significantly better performance and portability characteristics than using Python would have allowed for.

5.2 Pipeline Implementation

This subsection describes the system design realisation of each core pipeline component (σ, ϕ, π, ρ, ν) into working code, their algorithmic details, performance characteristics and trade-offs made.

5.2.1 Example Selection Engine (σ)

The example selection system is implemented using a multi-dimensional similarity scoring system that identifies the most relevant samples that were previously indexed for subsequent in-context learning. The component addresses the fundamental challenge of selecting contextually appropriate examples from large training corpora whilst balancing semantic relevance and structural compatibility.

5.2.1.1 Similarity Computation

The selection mechanism combines three distinct similarity measures implemented in the selection module in `natural-sampling::selection`. The `selection` algorithm computes semantic similarity through cosine distance of the question embedding. Structural SQL similarity is computed via measuring the cosine distance of the SQL embedding, and schema compatibility using WWL kernel distance from the `natural-graphs` component. The embeddings are computed using the Qwen3-Embedding model (4B) using the embedding implementation in `natural-sampling`.

5.2.1.2 Weighting Strategy and Performance Characteristics

As this algorithm is aware of three distinct ways to measure sample similarity (question, query, structural). The implementation employs empirically optimized weights defined as constants: 70% for semantic question similarity and 30% for SQL structural similarity. Additionally, the final scoring combines 70% sample-level similarity with 30% schema-level compatibility.

The selection algorithm yields a maximum of 32 candidates (`TOP_K = 32`) to limit computational overhead whilst maintaining selection quality. Vector similarity queries are performed using SQLite with the `sqlite-vec` extension, achieving sub-second retrieval times for vector databases exceeding 2,000,000 samples.

5.2.1.3 Implementation Architecture

Using the `Selector` struct the embedding computation and representation is encapsulated from the actual selection algorithm. Thus via `Selector::new` the consuming code can compute the embeddings and use them subsequently to run example selection.

The `selection` algorithm implementation follows a two-stage approach, split into initial candidate retrieval through vector search (using the `Vector` struct and SQLite) and subsequent reweighting using the above described weights and a precomputed WWL distance index from `natural-graphs`.

5.2.2 Schema Subsetting System (ϕ)

The schema subsetting system is implemented in `natural-inference::pipeline::subsetting` as `SchemaSubsetter` struct. The core algorithm is the `optimize` method (lines 21-47) which performs query validation to determine which tables are crucial for the query candidates provided.

5.2.2.1 Query Validation

Query validation employs a trial-and-error approach where for each query candidate and table in the schema, an in-memory SQLite database is created using `Connection::open_in_memory()`. Subsequently the connection is initialized using the schema with all tables in the schema except the current one.

Once the connection is ready, the `SchemaSubsetter` prepares every query candidate against this reduced schema. If preparation fails, this indicates the table removed is crucial for the query, thus it gets added to the list of crucial tables. This process is repeated for all query-table combinations to build a minimal schema containing only essential tables for the execution of the set of query candidates provided to the subsetting algorithm.

5.2.2.2 Performance and Trade-off Characteristics

As a new in-memory database for each table-query combination is created, this leads to $O(queries \times tables)$ complexity. For schemas with XXX+ tables and multiple query candidates, this results in XXX-XXXms processing time compared to 10-15ms for heuristic approaches.

However, this execution-based approach provides superior correctness guarantees since heuristic approaches cannot determine whether candidates will execute successfully in real database environments. Using a real SQLite in memory instance both the correctness of query candidates and the actually referenced set of tables can be known prior to actual query execution.

5.2.2.3 Pipeline Integration

The `SchemaSubsetter` is used prior to prompting the model using ICL to ensure that the model context is used efficiently and attention is given to the relevant parts of the schema. The `optimize` method returns a `Schema` object containing only the tables identified as crucial through execution testing which is in turn processed by the subsequent pipeline stages like generation and refinement.

5.2.3 Query Projection (π)

The ICL module is implementing the query projection algorithm described in section 4.2.3. It is implemented in the `ICLGenerator` struct and wraps the `SqlModel` struct from `natural-models`. Using llama.cpp it runs model inference using a prompt optimized for in-context-learning with OmniSQL.

5.2.3.1 In-Context Learning

The `ICLGenerator::generate` method implements few-shot prompting with relevance filtering. Relevance filtering refers to removing all selected samples with a similarity of less than 0.5 (this value was emperically derived from evaluations). Thus only semantically or structurally similar samples are actually provided to the model.

The prompt is constructed based on an adapted version of the OmniSQL format (H. Li et al., 2025): task overview, sql schema, filtered examples with similarity scores, explicit instructions, and the target question. The biggest differentiation to the prompt of H. Li et al. is the example section including similarity. For the actual SQL query presentation the code representation prompt format is used, inspired by DAIL-SQL (D. Gao et al., 2023). Every example includes the question, similarity score, and formatted SQL query for clarity.

5.2.3.2 Prompt Engineering Strategy

As outlined above the OmniSQL prompt was used as base for Natural's prompt together with a code representation prompt. Natural uses more explicit instructions (see **??**) compared to DAIL-SQL. Key differentiations include the inclusion of similarity scores to give the model the ability to weight the samples itself.

This prompt steers the model towards precision and chain-of-thought reasoning. It addresses apparent LLM issues like verbosity, over complexity or missing query constraints, as well as the hallucination and accuracy concerns identified in literature review section where LLMs generate plausible but incorrect SQL queries.

5.2.3.3 Model Integration and Performance

The `SqlModel` is wrapped around `llama-cpp-2` bindings and loaded globally at startup to avoid a 30-60 second initialization times per query. Using the `prompt` method from `natural-models` tokenization, inference, and decoding with configurable PromptParams for context size and generation limits is handled automatically.

5.2.3.4 Output Processing and Validation

As OmniSQL was finetuned to output its thoughts and predictions using markdown a markdown postprocessing module is needed, as well as a module to identify whether a SQL query is syntactically valid.

To extract possible candidates all generated responses are post-processed through a markdown parser (`pulldown-cmark`) which parses the model output and looks for the code-block fence characters ‘‘‘. Subsequently all candidates are processed in reversed order and the first full valid query is returned as potential candidate. The reversing of processing order is needed as the model outputs it's thoughts top-to-bottom as a markdown document with the most likely answer usually being output at the end. This approach ensures that a returned query candidates are executable and ensured to contain valid SQL. Thus Natural can aid the difficulty of generating perfect SQL queries by acknowledging limitations from large language models and implementing recovery and refinement mechanisms in the subsequent pipeline flow.

5.2.4 Self-Refinement Mechanism (ρ)

The self-refinement algorithm described in 4.2.4 is implemented in `natural-inference::pipeline::refinement`. This implementation corresponds to the ρ function, providing automated error correction through execution feedback and iterative improvement of generated SQL queries.

5.2.4.1 Error Correction Through Execution Feedback

The `Refinement::optimize` method takes a `RawQueryCandidate` and attempts to improve it through targeted prompting. The refinement prompt explicitly instructs the model to “spot any errors in the SQL query, correct them” and provides the original question, schema context, and the candidate query that needs improvement.

Contrary to the ICL generation which uses few-shot examples, the refinement process focuses on single-query self-correction with explicit error-fixing instructions.

5.2.4.2 Prompt Engineering for Correction

The refinement prompt uses a structured format similar to ICL generation but with key differences: it includes the original question as context, provides the candidate query that needs fixing, and uses explicit correction language (see **??**).

5.2.4.3 Model Resource Management

As both the ICL implementation and the refinement are ultimately generating SQL, they reuse the same underlying `SqlModel` from the shared pipeline context rather than loading separate models. This design choice significantly reduces memory requirements but potentially impacts refinement quality compared to having dedicated refinement models with different prompt conditioning. Due to the limited available hardware the effects of having a separate, distinc refinement model could not be verified. Y. Gao et al. have implemented multi model generation pipelines in 2025 and achieved promising results which.

5.2.4.4 Output Processing and Validation

Similar to ICL generation, refined responses are processed through the same `Markdown` parser to extract the predicted SQL. The system validates each extracted candidate by trying to parse it using `QueryCandidate::try_from` and returns the first syntactically valid refinement starting from the bottom. This ensures that refinement produces executable SQL whilst falling back gracefully if refinement fails.

5.2.4.5 Integration with Pipeline

The refinement module is integrated into the main pipeline after the ICL generator. It processes the `RawQueryCandidate` and provides a self-correction mechanism that improves overall pipeline robustness. Notably both the unrefined and the refined query candidates are kept in the candidate set for majority voting.

5.2.5 Consensus Voting System (ν)

The `natural-inference::pipeline::voting` implements the majority function ν – this implementation closely follows the design and algorithm outlined in the section 4.2.5. The function `voting` providing a result-based self-consensus mechanism which takes in the set of candidates that were predicted during the generation phase and returns the most likely query.

5.2.5.1 Result-Based Self-Consensus

H. Li et al. have shown in 2025 that self-consensus can significantly improve the accuracy of models that are already showing state-of-the-art performance. The result-based self-consensus mechanism is executing every query candidate, verifies that it works on the database, and loads it's results. By partitioning the available candidates into buckets based on their hashed result, queries can be deemed semantically equivalent if they end up in the same bucket. After every query has been exeuted and partitioned, the algorithm groups the buckets in a hash map `buckets` with the result hash as key and bucket as value.

5.2.5.2 Bucket Selection Heuristic

The heuristic method employed by the voting algorithm is steering the pipeline to agree with itself – if multiple generation attempts yielded the same result, these generation attempts are more likely to be accurate the others.

5.2.5.3 Error Handling and Fallback

Queries that fail execution are contained in an `errors` hash set rather than being included in voting buckets. If all candidates fail execution, the voting function fails with an error that contains all collected execution errors to provide diagnostic information to the calling side.

5.2.5.4 Limitations and Trade-offs

The result-based partitioning approach has limitations with regards to subtle errors that differ by few rows, as these would be treated as completely different result buckets. Furthermore executing every query candidate against the real database comes along with a significant performance penalty for expensive queries where the voting step might incur load onto the database, use significant amounts of memory for loading in all data into memory and slow down the voting as the entire result needs to be hashed. This system lacks model uncertainty calibration for SQL confidence scoring which could be used as another dimension for partitioning.

Another more advanced optimization could be to group by result-schema and row count, which still needs to execute the candidate queries partially but does not load the actual data from the database into memory. For the scope of this thesis and small real-world application scenarios the cost of this stage is negligible compared to the llm inference done in previous steps.

5.2.5.5 Integration with Pipeline

The voting mechanism represents the last step of the pipeline function and selects the final consensus from all generated and refined candidates, and ultimately returns the output of the NATURAL system.

5.3 Supporting Systems and Optimizations

The supporting systems of NATURAL are important tools to simplify the actual pipeline development, employing performance optimizations and enable rapid development of new approaches and hypotheses. The primary support systems are `Vector` – a vector database abstraction on top of SQLite, `wwl` – a library that implements Wasserstein-Weisfeiler-Leman kernels (Togninalli et al., 2019) and the `natural-sampling` module to compare and semantically search the embedding space.

5.3.1 Vector Database

The vector database implementation `Vector` is used both during sampling to construct the vector db and runtime to run the sample selection algorithm. Using SQLite and `sqlite-vec` to implement cosine search in text corpora of up to a few million samples dramatically simplified the development of NATURAL as no manual performance optimizations had to be implemnented. SQLite has a second architectural benefit besides relatively good performance, which is the portability characteristics of having all samples indexed in a single file on disk. This enables swapping out the respectively used samples by using another `.vector` file on disk. See A.2.1 for reference.

5.3.1.1 Database Schema

As the `.vector` file is still a regular SQLite database, alongside indexing tables using embeddings (`float[4096]` or `float[2048]`) regular database tables can be maintained to save data derived at sampling time. `Vector` hosts the database schema described in A.2.1.1.

Which maintains the precomputed distance indices and the database graphs for example selection. Thus at pipeline runtime the database indices and graph representations can be reconstructed through a cheap selection from `Vector` even though they are not indexed through embeddings.

The schema of the `samples` table largely enables the example selection at runtime, it maintains embeddings of the sql query and natural language question respectively. Thus after cosine similarity search the original query or question can be reconstructed as well as the graph layout of the underlying database can be accessed through the `database` table.

The `database` and `database_indices` table contain `JSON` columns respectively for serializing complex graph and graph index structures to the database.

5.3.2 Embedding

The embedding functionalities in `natural-sampling` abstract batch embedding of questions and SQL queries using the GPU so that the runtime and sampling code can efficiently compute and use embeddings.

5.3.2.1 Architecture

The primary type for embedding text is the `SemanticString` struct which holds strings alognside their embedding vectors. Furthermore this struct offers `embed_seq` and `embed_chunked` methods which can be used for batch computation of embeddings.

5.3.2.2 Batch Computation

During sampling, offloading data to the GPUs memory is an expensive operation. When sampling hundred thousands or millions of samples, incurring the IO roundtrip from CPU to GPU for computing a single embedding is a performance bottleneck that increases the sampling time to an order of days (eg, when sampling the SYNSQL) dataset. Thus batch processing can help to minimise the IO roundtrips needed between the CPU and GPU before computing embeddings and better utilise the available computing power.

Using the `embed_seq` method, a sequence of strings can be embedded and a sequence of `SemanticStrings` is returned. Furthermore through the `embed_chunked` method, chunks of unrelated data can be embedded in a single operation. The method maintains the original chunk layout which makes it possible to reconstruct the input semantic seamlessly. This optimization is used to embed the questions and sql queries alongside while maintaining clear separation in the embedding output.

```
INSERT DIAGRAM
```

5.3.3 Wasserstein-Weisfeiler-Leman Kernels

WWL kernels are used to construct the distance index $d \in \mathcal{D}$ for the σ function (see section 4.2.1) by computing schema distance through graph-based distance metrics. The underlying WWL implementation used by NATURAL is the reference implementation of WWL kernels from Togninalli et al., which is implemented in Python. Thus NATURAL can't directly use the WWL implementation as Rust and Python FFI is not directly possible. In order to buy into the Python libraries and ecosystem (eg, optimal transport) and utilize the existing implementation of the WWL kernels, writing a thin layer of Rust bindings around Togninalli et al.'s implementation was

deemed most sensible. Using `pyo3` calling Python code is made possible and let's NATURAL hook into existing ecosystems where needed. Due to the design laid out in section 4 the WWL implementation is only needed during sampling time and not required at runtime as all distances are precomputed and stored in a distance index, keeping the runtime portability characteristics of using Rust that were discussed above.

5.3.3.1 Rust-Python Integration Architecture

The WWL kernel library used by NATURAL is implemented using `pyo3` around the existing Python `wwl` library. The Rust bindings are usable through the `wwl` crate. The `WWLKernel` struct encapsulates the Python module reference and provides type-safe interfaces for both categorical and continuous propagation modes through respective methods. Using an existing graph library (`petgraph`) the complexity of the Rust bindings could be kept minimal, but required conversion from Rust's `petgraph` graph representation to Python's `igraph` graph representation.

5.3.3.2 Schema Graph Representation

The `natural-graphs` library implements the schema graph representation discussed in section 4.1.2.1 by parsing SQL statements using the `sqlparser` library and constructing an undirected `petgraph` graph instance. Nodes represent tables and columns and edges capture foreign key relationships and column ownership. Node labeling is used to encode schema constraints hierarchically as outlined in section 4.1.2.1.

5.3.3.3 Propagation Mechanisms

The `wwl` crate supports both categorical and continuous Weisfeiler-Leman propagation schemes. Categorical propagation operates on discrete node labels through iterative label refinement, suitable for constraint-based schema comparison. Continuous propagation utilizes node features encoded in matrices, enabling similarity computation based on quantitative schema properties like column cardinalities or data type frequencies (Togninalli et al., 2019). As NATURAL projects the database schema into a graph representation with categorical node labels, the categorical propagation scheme is used in `natural-graphs` to determine schema distance. The `wwl` bindings allow configuration of the kernel through the `KernelConfig` and `DistanceConfig` structs, which provide control over the iteration count (which defaults to 3), sinkhorn approximation settings etc.

5.3.3.4 Distance Computation and Caching

The pairwise Wasserstein distance is computed between the current database schema and each sample database during sampling phase. The distance index is a `HashMap` of a sample database name to it's distance to the current database NATURAL is running on. Caching of this computation is an efficient strategy to minimise runtime inference time as the distance index is static as long as database schemas are not mutated. NATURAL is storing JSON-serialized distance indices in `Vector` for fast distance lookups at runtime. This design prevents the computational cost of WWL distance computations from affecting inference, reducing runtime schema comparison to constant-time lookups for all previously indexed database combinations that were known during sampling time. Given that the set of sample database and target databases are usually fixed, the distance lookup becomes negligible in terms of computational cost.

5.3.3.5 Integration with Example Selection

The WWL kernel is integrated in the sampling phase, where distance indices are computed and cached. Furthermore during example selection (σ) the cached distance index is loaded from `Vector` for distance lookups.

The distance index is used for weighting the schema compatibility in the selection algorithm (see section 4.2.1) to balance structural similarity against question-level semantic relevance for improved in-context learning effectiveness.

5.4 Benchmarking Infrastructure

NATURAL's benchmarking infrastructure is implemented in `natural-benchmark`. This benchmarking infrastructure is enabling the development, verification and ultimately deployment of NATURAL pipeline. This section focuses on the respective benchmarking infrastructure development, performance optimizations that were required and engineering challenges encountered.

Whilst NATURAL's benchmarking infrastructure is not part of the core pipeline, it helps to develop and test hypotheses for extensions or design changes confidently, compare performance across pipeline versions, models and understand the benchmark performance.

5.4.1 Execution-Based Evaluation System

The `natural-benchmark` CLI evovles around the concept of executions which are single benchmark runs against a specific benchmark dataset (defaulting to SPIDER). An execution is the set of tests that have been executed using a version of NATURAL against a benchmark. The interface of `natural-benchmark` offers multiple options to create executions and subsequentlty understand NATURAL's performance:

1. Running new benchmarks via:
   `natural-benchmark new`
2. Continuing a halted execution via:
   `natural-benchmark continue -execution <id>`
3. Comparing performance on previous (partial) executions via:
   `natural-benchmark compare-to -previous <id>`
4. Computing statistics on past executions via:
   `natural-benchmark stats -execution <id>`

While benchmarking `natural-benchmark` maintains a local SQLite database maintaining a history of past performance, pipeline failures etc. for future introspection as well as comparison of approaches.

5.4.2 Cross-Dataset Validation

In order to measure the performance of NATURAL against multiple benchmarking datasets, the benchmarking and execution system in `natural-benchmark` must be generalize across one dataset. The benchmarking setup achieves this through a set of rust traits (eg, type characteristics) to model benchmarking datasets.

The traits `Benchmark`, `BenchmarkDatabase` and `BenchmarkTest` allow to abstract the file system layout of different evaluation datasets. Thus when implementing benchmark support against SPIDER the implementation absatracts the fact that SPIDER is using SQLite for test databases, and uses a JSON file to store the questions. The file system layout of SPIDER and BIRD is largely similar.

This abstracted benchmarking system (see A.2.1.2) allows integrating further benchmarks in the future (eg, SPIDER2 which relies on cloud databases like Snowflake) while maintaining the same ergonomics and tooling across benchmark datasets (eg, recording and indexing all test executions in a local database, recording pipeline logs etc).

5.4.3 Metric Computation

The benchmarking infrastructure computes the two metrics EXECUTION ACCURACY and EXACT MATCH found in prevalent benchmark leaderboards (eg, SPIDER and BIRD) through a Rust port of the Python reference implementation found in SPIDER. SPIDER determines the EXECUTION ACCURACY by comparing the result sets of two queries based on possible row and column permutations as well as optional order sensitivity. The EXACT MATCH metric is determined by exact equivalence of the ground truth query ω_{ground} and the candidate query ω.

5.4.4 Benchmarking Challenges

Two primary issues emerged while developing the benchmarking infrastructure for NATURAL:

While loading result sets from the provided test database, a faulty code path failed to parse query result with types other than `Text`. This casued empty result sets to be returned as query results from both the ground truth query and the candidate query, resulting in a false positive, skewing the EA metric significantly upwards.

Furthermore frequent pipeline failures of NATURAL made it hard to get compute reliable performance metrics as context window constraints, out of memory issues and CUDA issues caused a signficiant number of test cases to fail, making subsequently computed metrics unreliable as only a subset of the dataset was included in the results.

5.5 Engineering Challenges and Lessons Learned

The implementation approach of NATURAL highlighted significant challenges in terms of practicability that were independent of the algorithmic design and theoretical problems. While NATURAL turned out to demonstrate the approaches recommended in the system design section, the development process was significantly slowed down due to constrained hardware availability, technology stack decisions and research methodology. The overall system architecture turned out positive, but the timeline was noticably expanded.

This section reflects on the most impactful challenges that were encountered during the implementation of NATURAL and analyzes how hardware limitations influenced design decisions, confidence in the performance of NATURAL and the overall impact on development velocity. These insights are transferrable beyond NATURAL and are applicable to any production-grade implementation of research algorithms.

5.5.1 Hardware Constraints and Development Velocity

Likely the most significant circumstance that affected development velocity was the constrained access to high performance hardware. The hardware available during development time was the RTX 3090 with 24GB of VRAM. This limitation fundamentally shaped the development process and research methodology as the iteration speed was severly slowed down.

5.5.1.1 Benchmarking and Evaluation Bottlenecks

A full benchmark execution of NATURAL against the SPIDER test dataset required around 12-36 hours depending on the number of candidates generated and refinement algorithm used during the benchmark. This made continuous validation during development a non trivial task, requiring extensive evaluation phases between times of active development.

Testing new hypotheses, validating algorithmic changes, experimenting with the implementation approaches, comparing different configurations and sampling strategies therefore became a multi-week processes rather than the rapid iteration typically desired in research environments.

The prolonged evaluation cycles created a cascading effect on development velocity. Rather simple adjustments that would normally be validated within hours required days of computational time, effectively preventing what should have been normal experimentation. The hardware constraints forced a rather conservative approach to experimentation, where each change needed to be carefully considered before committing to multiple days of benchmarking.

5.5.1.2 Memory Management and CUDA Issues

Frequent CUDA out-of-memory (OOM) errors made the evaluation interpretation non trivial as accuracy scores became inaccurate due to pipeline failures. This persistet as a second development challenge throughout the implementation phase. These memory exhaustion issues occurred rather unpredictable during both development and benchmarking, depending on the database schema and question asked. As NATURAL uses a code representation prompt, the database schema is inlined into the prompt during inference. If NATURAL is executed against large databases, even with schema subsetting, the context window can be exceeded or the KV-Cache can exhaust the rest of the available GPU memory (typically around 2-4 gigabytes) after models have been loaded. This required extensive trial-and-error debugging to identify the root causes and implement workarounds.

5.5.2 Technology Stack Trade-offs

During the implementation of NATURAL, several critical technology stack decisions fundamentally impacted both development velocity and the final system characteristics. The two most consequential decisions were the choice of programming language and the model deployment strategy, each presenting distinct trade-offs between research efficiency and production readiness.

5.5.2.1 Programming Language: Rust vs Python

The decision to implement NATURAL in Rust rather than Python represents perhaps the most impactful architectural choice made during the implementation phase. Choosing Rust presented the opportunity to port the research system into production environments soon after, but came at the cost of having to port existing benchmarking infrastructure and not being able to exactly reuse the inference code used by OMNISQL. This

decision significantly increased implementation time, likely doubling the development effort compared to what would have been required in Python. Using Python would have provided immediate access to existing machine learning infrastructure and could have leveraged existing implementations by previous researchers directly in the implementation of NATURAL. The rather mature Python ecosystem for NLP and ML would have eliminated the need to develop abstractions for model inference, vector database operations, and benchmark evaluation.

However, the Rust implementation as is provides significant benefits for production deployment and system integration. The resulting system can be embedded directly into database systems such as PostgreSQL through extensions, providing a path toward true production deployment that would be impractical with Python. The ability to compile an entire NL2SQL system into a single, standalone binary offers superior portability characteristics.

5.5.2.2 Model Deployment: Local vs Cloud-Based Inference

Using local and open source models reemphasized the limited local hardware access as using a cloud provider like OpenAI or Anthropic for inference would have singificantly speeded up the benchmarking time. As a primary research goal of this thesis is to explore the open source capabilities of NL2SQL systems, choosing a proprietary inference service was deeemed unviable.

5.5.2.3 Integration Complexity and Ecosystem Maturity

The machine learning ecosystem of Rust proved to be singificantly less mature than anticipated, requiring extensive custom development and research for functionality that would be available off-the-shelf in prevalent Python packages. The integration with Python libraries for specialized components like the Wasserstein-Weisfeiler-Leman kernels required rather complex FFI code using `pyo3`, further adding architectural complexity and potential stability concerns.

Despite these challenges, the resulting architecture demonstrates that multi language approaches can be viable when different components have distinct requirements. The Python integration was isolated to the sampling phase, preserving the runtime portability characteristics of the core Rust implementation while still leveraging existing, specialized libraries where appropriate.

5.5.3 Lessons Learned and Future Recommendations

The implementation of NATURAL showed multiple practical challenges associated with researching and implementing NL2SQL systems. The approach used in this thesis highlights several areas of consideration for future reserch projects in this field.

5.5.3.1 Hardware Infrastructure Requirements

Research of modern NL2SQL systems induces steep hardware requirements onto researchers. Priority should be given to securing access to significantly powerful hardware and computing infrastructure for faster iteration cycles.

While the 24GB VRAM constraint of the RTX 3090, may be reasonable for most consumer grade applications, it proved insufficient for efficient iteration in this field, especially given the involement of large language models. Systems with 48GB+ of VRAM would have enabled the usage of larger, more capable models without the quantization compromises that may impact system accuracy. More modern hardware (even on the consumer grade level) would allow for a tremendous benchmarking speed up (eg, using a 5090 would have resulted in nearly 200% increase in speed).

More critically, the availability of multiple GPUs would have enabled parallel experimentation and benchmarking, which could transform multi-day iteration cycles into more manageable timeframes. The ability to run concurrent experiments with different configurations, models, or algorithmic approaches would dramatically accelerate the research and development process and enable more thorough experimental validation.

5.5.3.2 Technology Stack

The implications of choosing between Rust and Python underlines the importance of technology choices with research goals and time constraints. For pure research projects focused on algorithmic development and experimentation, Python's ecosystem and extensive existing implementations provide clear advantages that likely outweigh the deployment benefits of compiled languages like Rust.

However, for projects with a clear path to production deployments or those intended to bridge academic research with real-world applications, the additional development overhead of systems programming languages like Rust can be justified. Overall, hardware availability weighs significantly more than language choices when it comes to the implementation timeline.

Future implementations in this field should likely consider a two-staged approach: prototyping and algorithm validation in Python to leverage existing tools and accelerate iteration, followed by a production port in a systems language like Rust once the algorithmic approach is validated and stable.

5.5.3.3 Development Methodology Recommendations

The challenges encountered suggest several methodological improvements for future projects:

Implementing benchmarking and validation infrastructure early in the development process is crucial when doing algorithmic development to gain an understanding of the relative improvement or setback of different approaches.

Establishing performance baselines and using incremental validation procedures can help ensure that development effort is focused on improvements that in fact improve performance.

The long running benchmarks made it difficult to validate whether individual changes were beneficial, leading to uncertainty about the effectiveness of specific optimizations. In situations with highly constrained compute resources, a representative subset of prevalent benchmarking datasets like Spider and Bird may be defined to determine the relative improvement between approaches before executing a full benchmark run.

Finally, maintaining detailed logs and metrics throughout the development process showed to be crucial when unexpected errors occur. Identifying root causes of pipeline failures was rather trivial due to having an advanced benchmarking setup with log collection.

5.5.3.4 Research Impact and Production Readiness

Despite all challenges encountered in the implementation phase, the Rust implementation of Natural demonstrate that it is possible to port research algorithms from academic environments into production-ready systems. The resulting implementation provides a strong foundation for real-world deployments in database environments, which would be significantly more challenging with a Python-based research prototype.

This production readiness ensures that all algorithmic contributions can have practical impact beyond pure academic evaluation, potentially influencing how NL2SQL systems are deployed in real database environments. A simple recommendation between research velocity and production readiness can not be given as it ultimately depends on the intended impact and longevity of the research contributions and implementation.

6 Evaluation

The evaluation phase of NATURALbuilds upon the benchmarking infrastructure outlined in section 5.4. This section is evaluating benchmarking data gathered while running NATURALin different configurations on two prevalent benchmarking datasets: SPIDERand BIRD. A representative baseline is introduced, aiming to mirror the base-model performance. Subsequently an ablation study is performed to understand the performance impact of individual pipeline components.

6.1 Experimental Methodology

6.1.1 Test Environment

The tests are performed on a consumer-grade linux system, as research- or enterprise-grade hardware was not available. Thus only a subset of datasets and configurations have been measured, as a single benchmark executions on prevalent datasets takes 12-36 hours.

6.1.1.1 Hardware Configuration

The system used for benchmarking has a NVIDIA RTX 3090 GPU with 24GB VRAM available, an AMD 9990X3D CPU with 12-cores up to 5.5GHz, 64GB of RAM at 6400mt/s and 1TB of SSD storage at a read speed of 7450 MB/s and a write speed of 6900 MB/s.

6.1.1.2 Software Stack

The software stack running on the system used for benchmarking is NixOS (25.05) with a recent linux kernel version (6.17.7). CUDA 13.0 is being used for inference on GPUs and a rust compiler version is 1.91 for compiling NATURAL. The models used during benchmarking are OMNISQL 7B Q8_0 and QWEN3 EMBEDDING 8B Q4_K_M for generation and embedding respectively.

6.1.2 Benchmark Datasets

The benchmarking datasets used are the SPIDERand BIRDtestdatasets that are prevalently used for evaluation of NL2SQL systems. Comparable research has frequently published performance metrics (EXECUTION ACCURACY and EXACT MATCH) on these benchmarks which makes the performance of NATURAL comparable to other works.

6.1.2.1 Spider

SPIDER is likely the most prevalently used benchmark for NL2SQL systems in contemporary research. The test dataset is comprised of 10,181 questions and 5693 corresponding SQL queries spread across 200 databases spanning 138 domains. Questions in SPIDERare categorized by difficulty (easy, medium, hard, extra hard). Evaluations on SPIDERrefer to the test dataset.

6.1.2.2 Bird

BIRDis another widely used benchmark that is comprised of 12,751 question answer pairs spread across 95 databases in 37 domains. It aims to be more representative of real world scenarios with external knowledge requirements and is generally considered harder than SPIDER.

6.1.3 Evaluation Metrics

The chosen set of evaluation metrics used is a mixture of semantic metrics ($\mathbb{EA}$, $\mathbb{EM}$) and functional metrics ($\mathbb{ER}$, $\mathbb{CL}$) to gather a hollistic picture of system behavior and real world applicablility.

6.1.3.1 Execution Accuracy ($\mathbb{EA}$)

Execution Accuracy (or $\mathbb{EA}$) is the primary success metric of NL2SQL systems as it represents the semantic accuracy of the SQL queries produced NL2SQL systems. Execution accuracy is computed by determining whether the rows in the candidate results are a permutation of the rows returned by ground truth results.

For a query q_g with candidate results R_c and ground truth results R_g, where $R_c, R_g \subseteq \mathbb{V}^{n \times m}$ (sets of n rows with m columns of values $\mathbb{V}$), semantically accurate execution is defined as:

$$semanticeq(R_c, R_g, q_g) = \nVdash \left[\exists \pi \in \Pi_m : \begin{cases} R_c = \pi(R_g) & \text{if } ordered(q_g) \\ R_c \equiv_{\text{multiset}} \pi(R_g) & \text{otherwise} \end{cases} \right] \tag{16}$$

where Π_m is the set of valid column permutations (those preserving column value sets), $\pi(R_g)$ applies permutation π to reorder columns in each row of R_g, $\equiv_{\text{multiset}}$ denotes multiset equality (allowing row reordering) and $ordered(q_g)$ is determining if the ground truth query q_g contains an `ORDER BY` clause.

In order to compute the $\mathbb{EA}$on SPIDERand BIRDthe official evaluation suite is used respectively to ensure comparability across externally reported measurements and locally observed measurements.

6.1.3.2 Exact Match ($\mathbb{EM}$)

Exact Match (or $\mathbb{EM}$) measures the syntactic equivalence between candidate and ground truth SQL queries after normalization (ie, whitespace trimming, casing adjustments etc). Unlike execution accuracy which validates semantic correctness through result comparison, exact match determines whether the generated query structurally matches the reference query. $\mathbb{EM}$acts as a lower bound for execution accuracy as queries must be identical which is therefore more restrictive as semantically equivalent queries with different syntax (ie, using a different join order) are marked as incorrect.

6.1.3.3 Error Rate ($\mathbb{ER}$)

Error Rate (or $\mathbb{ER}$) describes the system reliability by measuring the frequency of SQL generation failures that prevent query execution. $\mathbb{ER}$is reported in failures per hundred queries and monitors system reliability (schema violations, syntax errors, out-of-memory errors etc).

6.1.3.4 Candidate Latency ($\mathbb{CL}$)

Candidate Latency (or $\mathbb{CL}$) measures the end-to-end execution time of a system from natural language input to SQL candidate output, reported in seconds. This metric captures the compound runtime of all pipeline components ($\sigma, \phi, \pi, \rho, \nu$) and reflects real-world system responsiveness. Latency increases with pipeline complexity, particularly when example selection and self refinement are applied, thus representing relative computational cost.

6.1.4 Baselines

To adequately determine the performance impact of methods and components applied in NATURALa baseline helps to measure relative performance gains or losses compared to existing systems. Therefore two sets of baseline systems are introduced: Internal baselines (measuring the system without crucial components) and external baselines (eg, proprietary system performance and raw model performance).

6.1.4.1 Internal Baselines

Introducing two internal baselines subsequently allows for a brief ablation study, where the impact of different pipeline configurations and different sampling datasets can be measured to isolate the contribution of each configuration. The *Baseline* configuration of NATURALis using only the inference logic (π) without all other pipeline components, representing the performance contribution of OMNISQL as closely as possible. Additonally the *Zero-Shot* configuration of NATURALis introduced and measures the performance of NATURALwith all pipeline components activated but without any examples available to use during in-context learning.

6.1.4.2 External Baselines

Using published results from other systems, NATURALcan be briefly compared against existing systems with similar capabilities highlighting relative performance improvements or losses. Notably external baselines largely rely on unverified data from other papers. Given that NATURALis largely based on OMNISQL, GPT-4 and OMNISQLare the primary external systems are taken into account as baselines.

6.1.5 Vector Databases

Three different vector databases are introduced per benchmark. These differ in the datasets using during sampling:

6.1.5.1 Synthetic

The *Synthetic* vector databases are sampled from the SynSQL-2.5m dataset introduced by H. Li et al. in 2025. The SynSQL-2.5m dataset is largely unrelated to Spider and Bird but covers a wide array of domains through its LLM-based synthetic data generation approach.

6.1.5.2 Train

The *Train* vector databases are sampled from the respective training splits from Spider and Bird (J. Li, Hui, Qu, et al., 2023; Yu, Zhang, et al., 2018). The train splits include similar style of SQL queries and similar domains but different databases and different questions.

6.1.5.3 Ground

The *Ground* vector databases are sampled from the respective splits used during evaluation (dev and test) from Spider and Bird (J. Li, Hui, Qu, et al., 2023; Yu, Zhang, et al., 2018). These allow Natural to reference "previously used" answers as examples during generation indicating the potential upper-limit of performance during a self-learning deployment which has access to past conversations.

6.2 Benchmark Results

This section is presenting the measured performance across all Natural configurations and benchmark datasets. Internal ablations are compared (Baseline, Zero-Shot, as well as different vector DBs) against external prior art (GPT-4, OmniSql). Subsequently performance characteristics and improvement trends are quantitatively and qalitatively analyzed.

6.2.1 Overview

The overall benchmarking results are presented in Table 1 and visualized in Figures 7 and 8, which displays the the performance for all system configurations across the Spider and Bird datasets, the execution accuracy ($\mathbb{EA}$), exact match ($\mathbb{EM}$), error rate ($\mathbb{ER}$) and candidate latency ($\mathbb{CL}$) metrics.

6.2.2 Baseline Performance Analysis

One notable insight during benchmarking was the apparent gap between the and the locally measured performance of OmniSql 7B (using GGUF with the F16 weights) and the reported accuracy metrics results from H. Li et al. (2025). As shown in Figure 11, the δ in performance between the local measurement and the official values ranges from 2.6% on Spider (dev) to 11.8% on Spider (test) and 28.1% on Bird (dev). The section performance gap section (6.4) analyzes this gap in detail.

The *Baseline* configuration of Natural further showed degredation of performance compared to the OmniSQL-7B-gguf system configuration. Using the *Baseline* configuration, Natural only uses the inference logic and prompt template for ICL (π) without other pipeline components, achieving 77.9% $\mathbb{EA}$ on Spider (dev), 77.0% on Spider (test), and 32.4% on Bird(dev). These results further fall below the published OmniSQL-7B performance of 81.6%, 89.8%, and 66.1% respectively, representing gaps of 3.7, 12.8, and 33.7 percentage points. This continued degredation of performance indicates that a misaligned prompt template (ie, divergent from training phase) harms performance if not paired with counter measures such as ICL, self-correction or majority voting. These performance numbers mark the *Baseline* configuration the worst performing variant of Natural on $\mathbb{EA}$ and $\mathbb{EM}$ across Spider (dev and test) and Bird (dev).

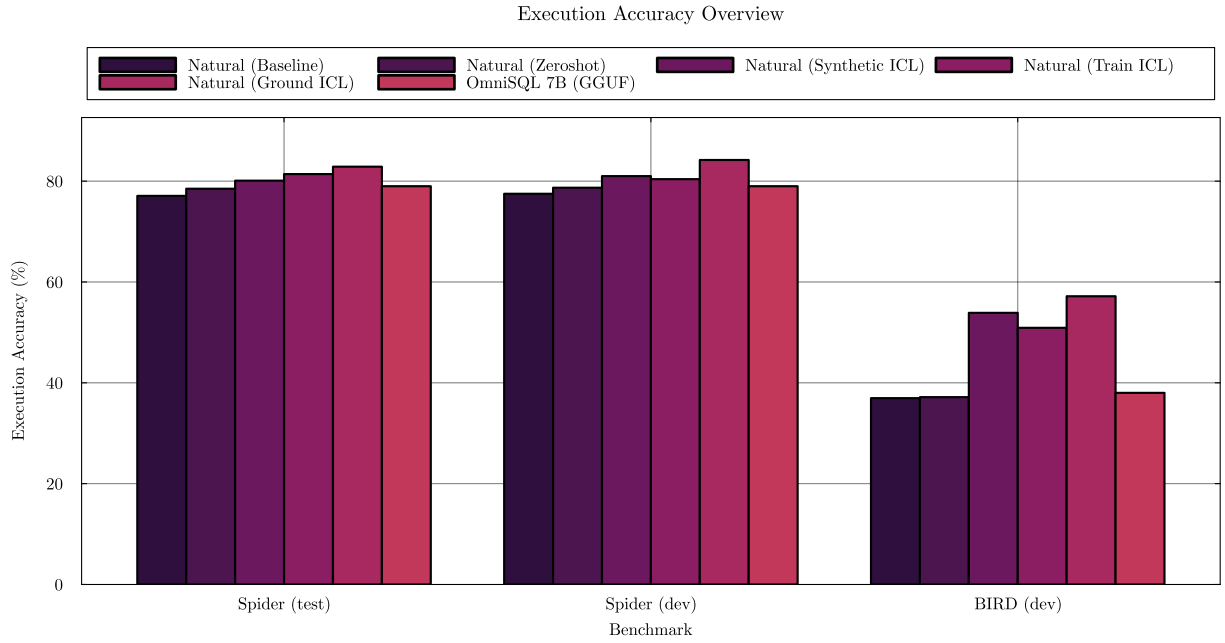

Figure 7: Execution accuracy overview across all systems and benchmarks. The *Syn* configuration achieves the highest verified performance on Spider (dev) at 81.0% and Bird (dev) at 53.8%, while *Train* leads on Spider (test) at 81.4%.

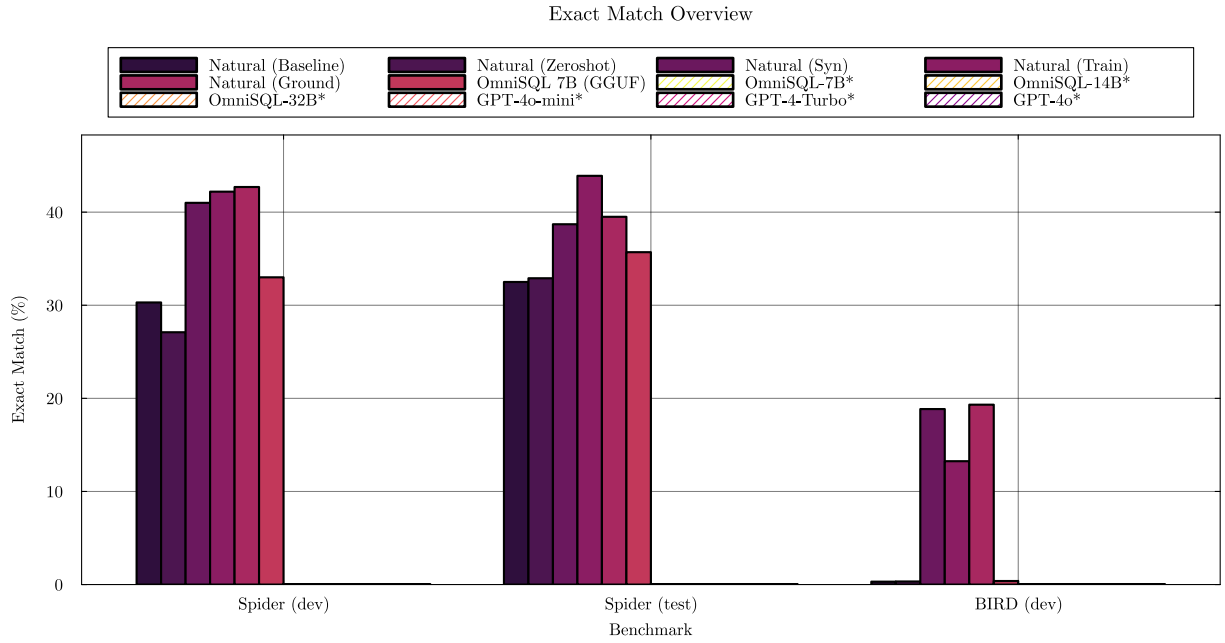

Figure 8: Exact match overview across all systems and benchmarks. ICL configurations demonstrate substantial improvements over baselines, with *Train* achieving 43.9% on Spider (test), the highest verified exact match rate across all benchmarks.

System	Spider (dev)				Spider (test)				BIRD (dev)			
	𝔼𝔸	𝔼𝕄	𝔼ℝ	ℂ𝕃	𝔼𝔸	𝔼𝕄	𝔼ℝ	ℂ𝕃	𝔼𝔸	𝔼𝕄	𝔼ℝ	ℂ𝕃
Closed-Source LLMs												
GPT-4o-mini*	71.0	-	-	-	83.7	-	-	-	61.5	-	?	-
GPT-4-Turbo*	72.2	-	-	-	84.2	-	-	-	63.6	-	?	-
GPT-4o*	70.7	-	-	-	84.9	-	-	-	64.0	-	?	-
Open-Source LLMs												
OmniSQL-7B*	81.6	-	-	-	89.8	-	-	-	66.1	-	?	-
OmniSQL-14B*	82.0	-	-	-	88.3	-	-	-	65.9	-	?	-
OmniSQL-32B*	80.9	-	-	-	89.8	-	-	-	67.0	-	?	-
OmniSQL-7B-gguf	79.0	33.0	0.1	**6.6s**	79.0	35.7	0.3	**6.7s**	38.0	0.03	0.8	**8.2s**
Pipelines												
Natural (Baseline)	77.9	30.3	**0.0**	7.3s	77.0	32.5	**0.0**	7.3s	32.4	0.03	**0.3**	9.3s
Natural (Zero-Shot)	78.7	27.1	0.2	14.0s	77.5	32.7	0.4	15.3s	34.0	0.03	0.4	18.7s
Natural (Syn)	**81.0**	41.2	0.3	16.3s	79.6	38.7	0.4	15.4s	**53.8**	**18.8**	0.2	43.7s
Natural (Train)	80.4	**42.2**	0.6	16.2s	**81.4**	**43.9**	0.3	15.8s	48.4	13.2	0.3	33.7s
Natural (Ground)**	84.2	42.7	0.7	16.1s	82.8	39.5	0.3	16.0s	55.8	19.3	0.2	43.8s

Table 1: Comprehensive benchmark results across all systems and datasets. 𝔼𝔸 and 𝔼𝕄 are reported as percentages. 𝔼ℝ is reported in failures per hundred queries and ℂ𝕃 is reported in seconds. Systems marked with * are external benchmarks with unverified results from published papers. Systems marked with ** had access to ground truth data during inference, illustrating the upper bound achievable. Bold values indicate best verified performance, excluding systems with access to ground truth. (H. Li et al., 2025)

6.2.3 Spider Results

The results of the Spider benchmarks demonstrate consistent performance characteristics of Natural and OmniSql across the development and test splits (detailed breakdown in Appendix A.3.1). The OmniSQL-7B-gguf system outperforms the two Natural baseline configurations *Baseline* and *Zero-Shot* in 𝔼𝔸 and 𝔼𝕄by 2.1% and 0.3% on the development split and by 2.0% and 1.5% on the test split respectively.

The full pipeline with the *Zero-Shot* configuration (all components active but no examples) hence only yields marginal improvements (+0.8pp on dev, +0.5pp on test) over *Baseline*, suggesting that schema subsetting and query refinement contribute minimally without example-based guidance. Notably, as demonstrated in Figure 9, the ℂ𝕃of *Zero-Shot* increases by 6.7s on dev and 8.0s on test over *Baseline*, doubling the system latency without a significant improvement in system performance.

Introducing vector database access to the system configurations demonstrates substantial improvement in performance. The two configurations without ground truth examples showed a significant improvement in both 𝔼𝔸 and 𝔼𝕄: *Syn* outperformed all other system variants and baselines with a accuracy improvement of 3.1% over *Baseline* and 2.0% over OmniSQL-7B-gguf on Spider (dev). On Spider (test) the *Train* configuration outperformed *Syn* by 1.8% in 𝔼𝔸. For 𝔼𝕄 both system configurations show a steep improvement over the three baselines with largest 𝔼𝕄 deltas being observed between *Zero-Shot* and *Train* with +15.1% on Spider (dev) and between *Baseline* and *Train* with +21.4% on Spider (test). These double digit improvements in accuracy point towards the conclusion that in-context learning is an effective mechanism for further improving the performance of already fine-tuned models. The performance of the *Ground* configuration shows the theoretical upper limits in 𝔼𝔸 achievable on top of OmniSql with example selection algorithms with 84.2% on Spider (dev) and 82.8% on Spider (train) (see A.3.1).

Error rates remain consistently low across all configurations (<0.7 failures per hundred queries), see 10, indicating that failures are predominantly semantic (incorrect results) rather than syntactic (malformed SQL), which validates the robustness of OmniSql's SQL generation capabilities.

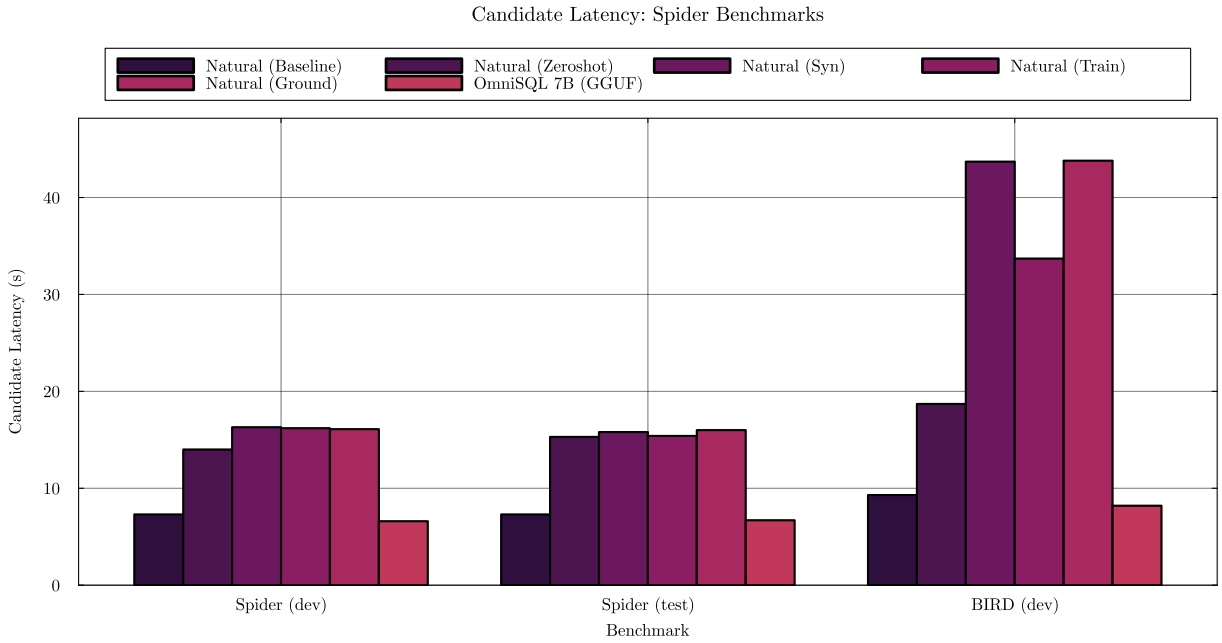

Figure 9: Candidate latency across system configurations on SPIDER benchmarks. The *Zero-Shot* configuration shows doubled latency (14.0s on dev, 15.3s on test) compared to *Baseline* (7.3s) without commensurate performance gains, while ICL configurations incur additional overhead from example retrieval.

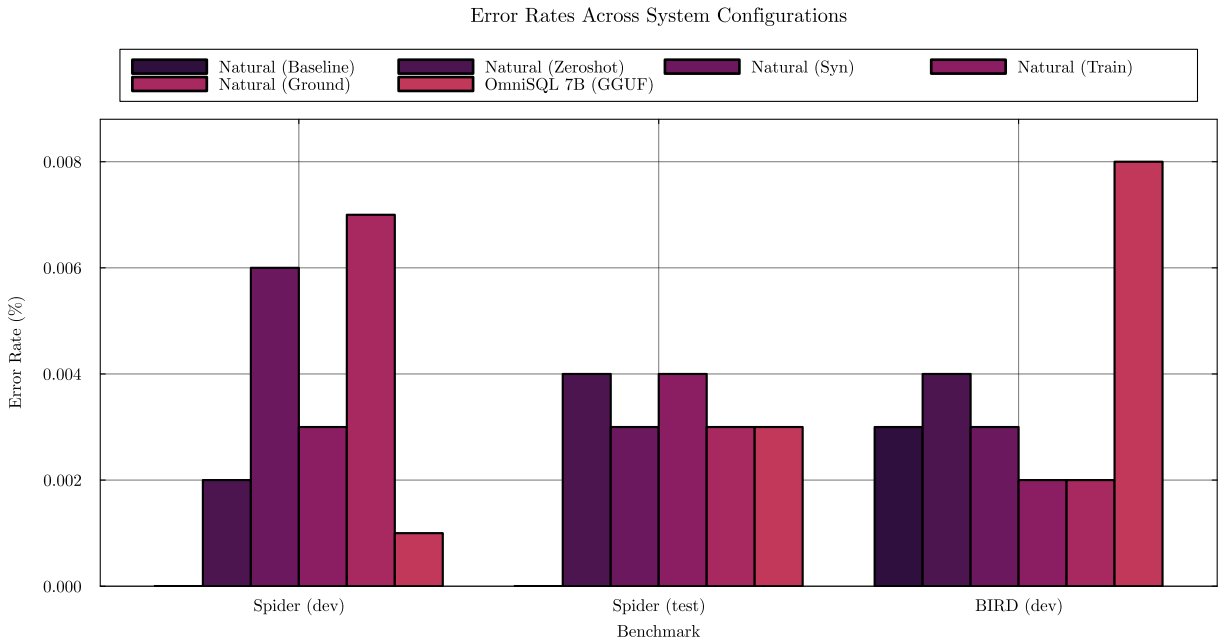

Figure 10: Error rates across system configurations. All configurations maintain error rates below 0.7 failures per hundred queries, indicating that failures are predominantly semantic (incorrect results) rather than syntactic or runtime related (malformed SQL, out of memory errors).

6.2.4 BIRD Results

Measurements on the BIRDbenchmark, which is designed to represent real-world scenarios with imperfect data, external knowledge requirements and hard to reason about database schemas yields significantly lower absolute performance numbers but shows fundamental improvements through ICL based systems. While the *Baseline* configuration of NATURAL only achieves 32.4% $\mathbb{EA}$ (less than half the accuracy than on SPIDER with 77.9% and 77.0% for dev and test splits respectively). This poor baseline performance confirms BIRDbeing a significantly more challenging benchmark. While the *Zero-Shot* configuration showns incremental improvements over the baseline with a relative improvement of +1.6pp in $\mathbb{EA}$ (34.0% absolute), the overall performance shows to be behind the reported numbers of OMNISQL (7B) and GPT-4 by H. Li et al. (2025) which were reported to had an $\mathbb{EA}$of 66.1% and 61.5% respectively. The locally reproduced measurement of OMNISQL (7B-gguf) yielded a drop in $\mathbb{EA}$ of -28.1pp compared to the officially reported $\mathbb{EA}$ yielding 38.0% in absolute accuracy.

Both the *Syn* and *Train* configurations of NATURAL outperform all baselines. *Syn* showns an improvement of +21.4pp over *Baseline* and +15.8pp over OMNISQL (7B-gguf), achieving 53.8% $\mathbb{EA}$, the best local measured system excluding *Ground*. *Train* yields improvements of +16.0pp over *Baseline* and +10.4pp over OMNISQL (7B-gguf).

This shows a stark contrast with SPIDER, where synthetic examples provided only +3.1pp improvement on dev and +2.6pp improvement on train. The effectiveness of synthetic examples on BIRD suggests that the impact of ICL increases with query and schema complexity. When queries involve complex joins, nested subqueries, and aggregations, even domain-agnostic examples provide crucial scaffolding for the generation. The *Train* configuration of NATURAL underperformed *Syn* by 5.4pp. This may indicate that BIRD's training examples, while domain-relevant, contain different structural patterns than the development set, whereas SYNSQL-2.5M contains a broader coverage of different database schema structures to reference from.

The other metrics $\mathbb{EM}$, $\mathbb{ER}$ and $\mathbb{CL}$ further show a stark difference between SPIDER and BIRD. Systems which mostly relied on OMNISQL (NATURAL*Baseline* and *Zero-Shot* and OMNISQL (7B-gguf)) showed the same 0.03% in $\mathbb{EM}$indicating a potential overfitting of OMNISQL to the SQL style used in SPIDER. Notably ICL based systems showed a significant improvement with +13.17pp for *Train* and +18.77pp for *Syn*. The $\mathbb{CL}$ metric showed a similar pattern to the SPIDER benchmarks with the $\mathbb{CL}$increasing with pipeline complexity from 9.3s for *Baseline* to 43.8s for *Ground*.

Lastly the *Ground* configuration establishes an upper bound of 55.8% $\mathbb{EA}$, representing a +23.4pp improvement over *Baseline* (with a +72% relative gain). Despite these substantial relative improvements within the recorded benchmarks a significant gap of -10.3pp and -8.2pp remains when compared to the published OMNISQL (7B) performance with 66.1% and GPT-4o performance with 64.0%. This stark difference between local replication and officially reported performance metrics is yet unclarified.

6.3 Performance Characteristics

To assess the performance of NL2SQL systems hollistically metrics beyond semantic accuracy have to be taken into account. The two performance characteristics $\mathbb{ER}$ and $\mathbb{CL}$ provide insight into system reliability and their real-world applicability. These functional metrics complement accuracy by revealing operational constraints (such as hardware requirements or reliability) that affect deployment feasibility.

6.3.1 Error Rate Analysis

The error rate results demonstrate great reliability across all benchmarked systems and datasets. The error rates remain below 0.7% (fewer than 7 failures per 1000 queries) across all measured systems, with the *Baseline* achieving 0.0% error rate on SPIDER (dev) and 0.3% on BIRD (dev). Even NATURAL configurations with self-refinement, query parsing and ICL maintain low error rates: *Ground* shows an $\mathbb{ER}$ of 0.7% on SPIDER (dev), 0.3% on SPIDER (test) and 0.2% on BIRD (dev).

This consistency indicates that the OMNISQL model generates syntactically valid, executable SQL with a high reliability, and that pipeline components do not introduce significant failure modes.

Furthermore these low error rates reveal an important characteristic of OMNISQL and NATURAL: Errors of these NL2SQL systems are predominantly semantic (ie, wrong results) rather than syntactic (eg, malformed SQL or schema violations) or runtime-related.

This distinction matters for production deployments, as semantic errors are significantly harder to validate and catch than hard system failures. Semantic accuracy might be further improved through user feedback which can feedback into ICL.

6.3.2 Latency Analysis

Measuring the $\mathbb{CL}$ reveals the computational trade-offs that come with the the NATURAL pipeline architecture. The *Baseline* configuration achieves minimal candidate latency (6.6s on SPIDER dev, 6.7s on test, 8.2s on BIRD dev), representing the time required for schema subsetting and SQL generation without example retrieval or refinement.

The *Zero-Shot* pipeline configuration approximately doubles the end to end latency to (14.0s, 15.3s, 18.7s respectively) highlighting that the schema subsetting, self-refinement and voting components contribute to a significant increase in computational cost. Configurations with example selection and in-context learning show further latency increases. The *Train* and *Syn* configurations achieve 15s-17s $\mathbb{CL}$ on SPIDER and 33s-44s on BIRD, with the *Syn* configuration yielding the highest latency (43.7s on BIRD).

Overall OMNISQL (7B-gguf) consistently showed the lowest $\mathbb{CL}$ values with (6.6s, 6.7s and 8.2) for SPIDER (dev and test) and BIRD respectively. This is unsurprising as the OMNISQL (7B-gguf) system provides a realistic estimation for the π component of NATURAL.

These characteristics can inform potential hardware requirements or deployment decisions. For interactive applications sub-second response times of these models are recommended. NATURAL clearly exceeds these thresholds by a factor of up to 45 on complex databases and questions (eg, BIRD). Thus either significantly better hardware is required, a simpler configuration needs to be used or further performance optimizations need to be introduced.

6.4 Performance Gap

During baseline validations of the foundational models a critical performance discrepancy emerged (see 11) between the locally measured performance of OMNISQL 7B and the published performance by H. Li et al.. This section systematically breaks down the magnitude, potential causes and implications of this gap.

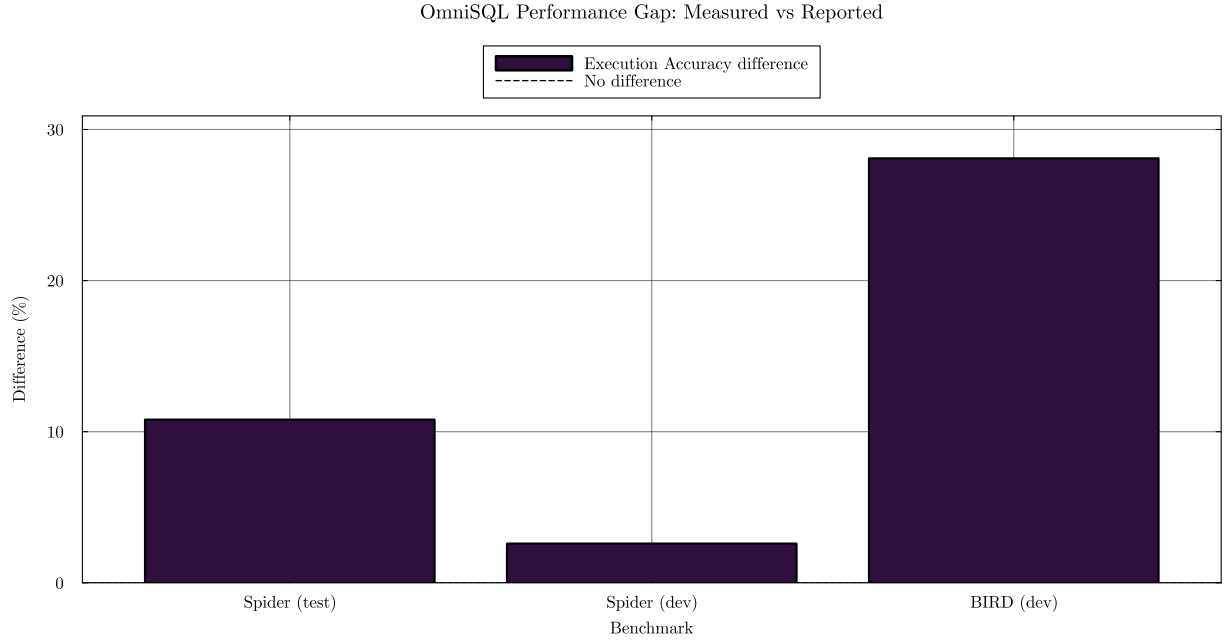

Figure 11: Performance gap between measured OmniSQL-7B-gguf and reported OmniSQL-7B results across benchmarks. The δ ranges from 2.6% on SPIDER (dev) to 28.1% on BIRD (dev), highlighting differences between quantized local deployment and published full-precision results.

6.4.1 Magnitude and Characterization

Table 2 presents the gap between the measured performance discrepancies across all evaluated benchmarks for the OMNISQL 7B model. It is important to note that a GGUF variant was used for local measurements.

The resulting gap varies significantly by dataset, ranging from only 2.6% on Spider (dev) up to 28.1% on Bird (dev).

Benchmark	Published (𝔼𝔸)	Measured (𝔼𝔸)	Gap (Δ)
Spider (dev)	81.6%	79.0%	-2.6pp
Spider (test)	89.8%	79.0%	-10.8pp
Bird (dev)	66.1%	38.0%	-28.1pp

Table 2: Performance gap between published OmniSql 7B results (H. Li et al., 2025) and local measurements using OmniSQL-7B-gguf (Q8_0 quantization). Negative values indicate underperformance relative to published baselines.

The uneven distribution of the measure performance gaps across the three benchmarks suggests a systematic cause rather than random variation. The gap correlates strongly with benchmark complexity: Spider (dev), the simplest of the three benchmarks with well-structured schemas and clean data is showing only a minimal performance degredation of -2.6pp. Spider (test) which is not included in any training dataset, using more diverse domains and more complex queries is showing a moderate degreation of -10.8pp. Bird (dev) which is the most complex benchmark of the three, featuring larger schemas, dirty data and external knowledge is showing an extreme degredation of -28.1pp.

This is indicating a performance drop correlating with task difficulty, suggesting that the locally measured model had impaired reasoning capabilities, difficulties handling larger schemas or handling external knowledge requirements.

6.4.2 Potential Contributing Factors

Several factors may contribute to this stark gap between the two systems. This subsection is analzing potential causes using available evidence.

6.4.2.1 Model Quantization and Format Differences

Likely the most apparent difference between the published performance metrics and the local measurements for OmniSql 7B is the model format and precision. The OmniSQL-7B-gguf system uses the F16 weights but transformed into the GGUF model format while published results presumably use full-precision models in the safetensors formats. Prior research has shown that bit-wise quantization can degrade model accuracy, particularly on complex reasoning tasks on multi-billion paramater models (Dettmers, Lewis, Belkada, & Zettlemoyer, 2022) and it is yet unclear whether the mere transformation into GGUF harms performance for this model series. Second, the exact hyperparameters used during evaluation likely differed between H. Li et al. (2025) and local benchmarks which may influence the model accuracy in unforseen ways. Unless the exact same inference environment, quantization and hyperparameters the performance difference might be caused by a multitude of subtle differences. Finally, dataset version differences or preprocessing variations could contribute to divergent results.

The Δ of 2.6% on Spider (dev) is well within the expected format-induced degradation ranges, which initially suggests that the model format alone could explain the simple benchmark discrepancy. However, the widening performance gaps with increasing benchmark complexity (up to 28.1% on Bird) substantially exceeed the typical effects of model format conversions indicating that other factors might contribtue to the observed performance loss.

6.4.2.2 Inference Implementation Differences

H. Li et al. likely used `vllm` as their inference engine, while local measurements of OmniSql 7B used Rust FFI bindings to llama.cpp. These engines implement the same mathematical operations but may differ in multiple dimensions. The engines might differ in implementation when it comes to numeric operations, the exact attention implementation, sampling algorithms and KV-cache management. To keep further inference environment divergent as small as possible, similar (where possible) sampling configurations, prompts and hyperparameters where used.

6.4.2.3 Hyperparameter and Configuration Uncertainty

While the published OmniSql evaluation does not exactly describe all its inference hyperparameters, some were noted in example code but uncertainty remains about the concrete configuration. Critical parameters that can widely influence model performance include temperature, top-p sampling, maximum generation length, context window size, repetition penalties and stop sequences.

Local measurements used a context size of 32768 and only greedy sampling with standard stop sequences. Inferring from the results published by H. Li et al. greedy sampling was the primary sampling method during evaluation alongside majority voting. Overall a difference in hyperparameters and inference engine configuration is possible, but is concluded to be unlikely given that greedy sampling as a standard and very straightforward algorithm.

6.4.2.4 Prompt Template Alignment

While a prompt drift could cause significant performance differences the exact prompt as supplied by H. Li et al. was used. Subtle differences in pharsing, layout or presentation could yield widely diverging results. It is likely that OmniSql 7B was fine-tuned using a specific prompt format, and deviation from this format can significantly degrade performance.

The local benchmarking implementation used the published prompt template but likely had a slightly diverging code representation implementation than the authors of OmniSql. The code representation implementation used by the benchmarking harness is based around a the `sqlparser` crate in Rust which includes an implementation for human readable formatting of SQL queries. While these modifications are subtle, they may inadvertently misalign with the model's learned expectations on the code representation format, confusing the generation process.

Further evidence for a possible prompt drift comes from the *Baseline* configuration results. Even when using only the inference logic (π) without additional pipeline components, Natural achieved 77.9% EA on Spider (dev) compared to OmniSQL-7B-gguf's 79.0%—a 1.1pp degradation despite using the same model and quantization. This suggests that even OmniSqlis noticeably sensitive to prompt template differences affect performance.

The prompt template used in the local benchmarks for Naturalis shown in Appendix A.1.1 and for OmniSqlin Appendix A.1.3. Natural's prompt deviated from the presumed original format through the addition of similarity scores for in-context learning examples, more explicit instructions and the likely modified schema representation.

6.4.3 Implications for Evaluation Validity

Due to time and resource constraints no exhaustive systematic evaluation of different quantization schemes, inference engines and prompt templates was performed. This singificantly limits the ability to draw definite conclusions on the exact root cause for the performance drift between the published performance numbers and the locally measured ones.

6.4.3.1 Impact on Absolute Performance Claims

The substantial gap between the published vs reproduced baselines undermines the confidence in *absolute* performance gains. The *relative* performance gains are excellent (eg, *Syn* achieving 53.8% vs. OmniSql 7B-gguf achieving 38.0% on Bird) but given that no clear absolute improvement could be measured (eg, *Syn* achieving 53.8% vs OmniSql 7B achieving 66.1% on Bird). Furthermore the comparisons with other external systems become unclear if local measurements face a significant gap in compared to the published performance numbers.

Thus this thesis avoids strong claims about absolute performance levels and instead focuses on the relative improvements achieved on local NL2SQL systems which is conistent with the evaluation environment and results.

6.4.3.2 Reliability of Relative Comparisons

Most notably, the performance gap between OmniSql 7B and OmniSql 7B-gguf does not infect the validity results measured between different variants of Naturaland between Naturaland OmniSql. All Natural configurations and OmniSql 7B-gguf have been evaluated using the same evaluation framework and are therefore strictly comparable. Natural showed strong performance improvements across increasingly complex

pipeline configurations. *Syn* showed an improvement of +21.4pp pver *Baseline* on BIRDwhich was reliably measurable regardless of the absolute baseline uncertainty.

Furthermore component ablations in the evaluation provid valid evidence for understanding the relative contributions of different pipeline components (minimal gain without ICL vs. substantial gain with ICL). Therefore the conclusion holds true that example selection is the most impactful component and that components exhibit synergistic rather than additive effects remain valid although the conclusion is only backed by same-environment evaluations instead of cross-study evaluations.

6.4.3.3 Broader Implications for LLM-based Research

The challenges encountered in this thesis reflect broader, systemic issues in LLM-based NL2SQL research. Without accessibility to the same hardware, models, specifications and evaluation frameworks a reliable cross-study comparison becomes close to unachievable. Intra-study comparisons remain valid but require significant time and hardware resources to achieve. This suggests that the NL2SQL research community would suggest from standardization of inference frameworks, comprehensive documentation of all hyperparameters and configurations used as well as a release of exact prompts, preprocessing code and scripts used during evaluation. Furthermore could multiple independent reproductions increase the confidence in claimed results.

6.4.4 Recommendations for Future Work

Based on this analysis several recommendations for further research in this area can be made:

6.4.4.1 Systematic Quantization Study

A dedicated investigation of the impact of model quantization on NL2SQL performance could provide guidance for better understanding the quantization to performance tradeoffs. This would allow for emperical decision making when deploying choosing model sizes and quantization into resource constrained environments. Prior research exists on the general impact of quantization on model performance, but having concrete datapoints specific to NL2SQL would clarify the observed behavior (Dettmers et al., 2022). This study should evaluate state of the art NL2SQL systems across multiple quantization levels (F16, Q8, Q6, Q4 and Q3) on all major NL2SQL benchmarks like SPIDER, SPIDER 2.0 and BIRD measuring both the accuracy and latency implications of different configurations. Furthermore an analysis could reveal correlation in performance behavior between quantization levels and query and schema complexity.

6.4.4.2 Cross-Framework Validation

To isolate inference engine effects and subtle differences implied through different model formats the same model should be evaluated across multiple frameworks. This study could include PyTorch and Transformers, llama.cpp, vLLM and ONNX and evaluate state of the art NL2SQL models in varying sizes against prevalent benchmarks. Conistent results across the frameworks would further increase the confidence in reproducibility of results where descrepancies would highlight discrepancies in implementation-specific details of algorithms and prompting.

6.4.4.3 Schema and Example Presentation

While research on schema and example presentation exist within the NL2SQL community it is not exhaustive and includes evaluations of different classes of schema and query presentation (D. Gao et al., 2023; Y. Gao et al., 2025). Further research in this field could highlight differences introduced through subtle changes in presentation of examples or schemas and recommend mitigation techniques like identifying the most performant presentation formats or fine-tuning base models to a specific presentation format.

6.4.5 Summary

Overall the gap of 28.1pp on BIRD between the published OMNISQL 7B results (66.1% $\mathbb{EA}$) and local measurements (38.0% $\mathbb{EA}$) represents a significant challenge for the absolute performance interpretation from this evaluation. While strong incremental performance gains could be observed (eg, *Syn* achieving 53.8% vs. OMNISQL 7B-gguf achieving 38.0% on BIRD) by applying the techniques outlined in this thesis (schema aware example selection, self-refinement, majority voting) no clear statement can be made on the absolute performance

of NATURAL. Multiple factors are possibly contributing to the observed gap which include model quantization, inference implementation, hyperparameter uncertainty, prompt template drifting and potential dataset or evaluation script version differences.

Despite these limitations, the gap does *not* invalidate the core contributions of this thesis. The relative comparisons within the consistent local evaluation environment remain reliable, therefore enabling valid, but thesis-local conclusions.

7 Discussion

This section is linking the evaluation results back to the research questions of this thesis which where outlined in the introduction section. It is discussing the findings of this thesis and deriving recommendations for actions and future research work in the NL2SQL community.

7.1 Summary of Results

The evaluation of NATURAL demonstrated the capabilities of open-source LLMs and showed that they can achieve competitive performance on prevalent structured NL2SQL benchmarks such as SPIDER and BIRD. Nonetheless significant challlenges remain for an enterprise-scale adoption of NL2SQL systems. NATURAL achieves up to 81.0% on the SPIDER development split using the *Syn* configuration and 81.4% on the test split using the *Train* configuration which represents a +3.1pp and +4.4pp improvement over the 77.9-77.0% performance of OMNISQL-7B respectively (Table 1).

On the complex BIRDbenchmark, the *Syn* configuration reaches a 53.8% $\mathbb{EA}$which represents a +21.4pp improvement over the 38.0% achieved by the *Baseline* configuration — representing a 66% relative gain in execution accuracy. These results position NATURALas competitive with the locally-reproduced OMNISQL-7B-gguf baseline (79.0% on SPIDER (dev), 79.0% on SPIDER (test), 38.0% on BIRD) while remaining below published accuracy scores of closed-source model results (GPT-4o: 84.9% SPIDER, 64.0% BIRD).

The ablation study performed during evaluation reveals that pipeline components tend to interact in synergy rather than additively, with in-context learning and the example selection algorithm of NATURALserving as the crucial differentiator for performance. The *Zero-Shot* configuration which uses schema subsetting, self-refinement, and majority voting but does not apply example selection achieves only a +0.8pp improvement on SPIDERdespite doubling the candidate latency from 7.3s to 14.0s. NATURAL configurations with active ICL and example selection yield up to +2.3pp on SPIDER dev, +3.9pp on SPIDER test and and +19.8pp on BIRDcompared to *Zero-Shot*. This shows that both schema subsetting and self-refinement techniques provide minimal benefit without in-context learning and example selection. Further ablation would help to understand the potential cross effects of components and the implications of disabling pipeline components like schema-subsetting, self-refinement and majority voting. Furthermore increasing the pipeline iterations from 1 to k would yield insights into self balacing and self correction across iterations of NATURAL.

A comparison of different example sources shows that structurally diverse synthetic examples are not necessarily performing worse than examples from the same dataset using the NATURAL example selection algorithm. Synthetic examples outperformed domain-similar training examples on BIRDby 5.4pp (53.8% vs. 48.4%), while perfect example selection yielded 55.8% accuracy and thus adds +2.0pp over schema-aware selection. This indicates diminishing returns of in-context learning from increasingly sophisticated example strategies.

Observed performance characteristics reveal a complex tradeoff for real-world applications. While all NL2SQL system that were evaluated maintain remarkably low error rates of less than 0.7 failures per 1000 queries (see Figure 10) errors are therefore becoming predominately semantic (ie. incorrect results). Furthermore candidate latency ranges from 7.3s for NATURAL (*Baseline*) to up to 16.3s for ICL configurations on SPIDERand 43.8s on BIRD. This yields NATURALviable for experimental, analytical workflows and asynchronous report generation but unviable for interactive systems like chatbot applications. While performance optimizations could be applied the hardware used is the primary bottleneck for the speed that is achievable as the majority of $\mathbb{CL}$is attributable to inference.

Lastly the $\mathbb{EA}$ gap of 28.1pp between the published OMNISQL-7B results (66.1% on BIRD) and local measurements (38.0%) raise concerns to the reliability of measurements and highlight systemic challenges in LLM-based research. The analysis of this gap concludes that it is likely attributable to quantization differences, inference engine variations and prompt template / code formatting sensitivity. Despite this remaining uncertainty in absolute performance comparisons, clear relative improvements from pipeline composition are observable and reproducible within a controlled environment.

7.2 Answering the Research Questions

To conclude this thesis, the research questions that were outlined in section 1.3 are answered.

7.2.1 Research Question 1: To what extent can open-source LLMs achieve competitive NL2SQL performance through pipeline composition and optimization?

The capabilities of open-source NL2SQL models achieve *partially competitive* performance on prevalent, structured NL2SQL benchmarks. Especially through the composition of different algorithms the performance can be significantly improved. The evaluation performed in this thesis demonstrates that a combination of in-context learning, schema subsetting, self-refinement and majority voting boosts the competitiveness of fine-tuned base models. Nonetheless significant gaps in execution accuracy remain for enterprise-level real-world deployments.

NATURALachieves up to 81.0% 𝔼𝔸 on the SPIDER (dev) benchmark using the *Syn* configuration (examples selected from the SynSQL-2.5m dataset, Table 1). This represents a +3.1pp increase improvement over *Baseline* configuration and a +2.0pp increase over the baseline model performance of OMNISQL-7B-gguf. On the test set of SPIDER the *Train* configuration is outperforming the *Syn* configuration by +1.8pp with an 𝔼𝔸 of 81.4%. *Train* is further showing an +2.4pp increase over OMNISQL-7B-gguf. On the more complex BIRD benchmark *Syn* outperforms *Train* by +5.4pp in 𝔼𝔸 and OMNISQL 7B-gguf by +15.8pp. However, on all three benchmarks noticable performance gap ranging from 2.6pp to 28.1pp was observed which reduces the trust in the measured results.

This yields an overall unclear picture, with NATURAL showing strong performance gains over the locally measured baseline model performance of OMNISQLbut no absolute gain over the published results for OMNISQLand other external systems by H. Li et al.. For definite conclusions on the competitiveness of local LLM-based NL2SQL systems more research is needed.

7.2.1.1 Pipeline Composition Effectiveness

Notably, the pipeline composition had notable impacts on the performance on NL2SQL benchmarks. The effectiveness of more complex pipeline configurations scaled drastically with database and benchmark complexity. On simpler benchmarks like SPIDER only marginal, single digit accuracy gains (+2.0pp and +2.4pp) could be observed while approximately doubling the candidate latency (from 6.6s and 6.7s to 16.3 and 15.4s).

On BIRD, the difference between the evaluated configurations of NATURALis significant. *Baseline* achieved only about 32.4% 𝔼𝔸 while *Syn* achieved 53.8% 𝔼𝔸 , representing a 66% relative gain over baseline performance through an altered pipeline composition.

The ablation study shows that the effectiveness of a pipeline configuration heavily depends on in-context learning and example selection, as the *Zero-Shot* configuration showed only marginal gains (+0.8pp on SPIDER (dev), +0.5pp on SPIDER (test) and +1.6pp on BIRD) over the *Baseline* configuration while effectively doubling the candidate latency from 7.3s to 14.0s on SPIDER (dev), 7.3s to 15.3s on SPIDER (test) and 9.3s to 18.7s on BIRD. This data is showing that a naive composition strategy of simply enabling all available algorithms is not cost effective and might produce counter productive returns on user exerpience as the latency increases are significant.

Lastly the *Ground* configuration of NATURAL establishes a theoretical performance ceiling using perfect data during example selection and consistently performed best-in class during local measurements. *Ground* achieved 84.2% on SPIDER (dev), 82.8% on SPIDER (test) and 55.8% on BIRD, representing a relative gain to the next-best performing configuration of NATURALof +3.4pp, 1.4pp and +2.0pp respectively. This shows that in-context learning and example selection strategies have their limits and are already performing close to them. Further improvements would likely require better foundation models or a different architecture alltogether. A visual comparison across the different configurations (Figure 7) confirms that in-context-learning-based approaches (*Syn*, *Train*, *Ground*) outperform other configurations with diminishing returns from increasingly sophisticated example selection strategies.

7.2.1.2 Comparison to Closed-Source Baselines

As no direct local comparison of NATURAL against closed-source models has been performed, no definite answer can be given to the performance gap between local NL2SQL systems and external, closed-source systems. The published results by H. Li et al. in 2025 report GPT-4o achieving 84.9% execution accuracy on SPIDERtest and 64.0% on BIRDdevelopment (H. Li et al., 2025). Compared to local measurements of the best-in class configuration of NATURALcompared to GPT-4o shows a performance Δ of +10.3pp on SPIDER (dev), -3.5pp on SPIDER (test) and -10.2pp on BIRD. Generally this suggests competitive but not superior performance, although it must be interpreted cautiously as the performance gap between the results reported by H. Li et al. (2025) and local measurements are non trivial. Given the uncertainties in measurements this thesis can't

claim a definite competitiveness compared to closed-source models. Further evaluations are needed to clarify the performance characteristics of all measured systems. Nonetheless, NATURAL presents a portable, at least partially competitive, alternative to closed-source systems.

7.2.1.3 Accuracy Ceiling on Consumer Hardware

As NATURAL was developed and evaluated on consumer hardware, the hardware constraints induced by it imply a ceiling of possible model size and speed characteristics. The RTX 3090 has 24GB VRAM and can theoretically accomodate OMNISQL 7B and OMNISQL 14B but given that NATURAL is using additional embedding models and leaves 2-4GB headroom for KV-caches etc. to prevent out-of-memory crashes. Therefore only the OMNISQL 7B model (15GB without quantization) is supported unless heavier quantization formats are used.

This makes NATURALaccessible to consumers with high-end consumer-grade hardware ($1,500 as of 2024-2025) without requiring enterprise-level hardware. The possible performance impact of using larger foundation models such as 14B and 32B was not evaluated. For enterprise real-world scenarios larger base models could be used if the available hardware allows for it. The evaluation performed in this thesis suggests that with the selected model sizes and the available hardware NATURAL has a theoretical execution accuracy ceiling of 84.2% and 82.8% on SPIDER (dev and test) and 55.8% on BIRD. This marks an upper bound achievable with the 7B models and the algorithms used. Therfore the foundation model capabilities emerge as the limiting factor rather unless an entirely different pipeline architecture such as agentic usage of LLMs is used. Nonetheless scaling the hardware and parameter sizes to 14B or 32B would likely yield significant gains in execution accuracy.

The latency characteristics of NATURAL remains in the practical range for non-interactive workflows but poses a significant challenge for local, interactive applications. While the *Baseline* configuration achieves 7.3s $\mathbb{CL}$, the optimal configurations of NATURAL (*Syn* and *Train*) only achieve a $\mathbb{CL}$ of 15.4-16.3s (Table 1). This makes NATURAL viable for exploratory analytics and report generation but not for interactive environments such as SQL learning or interactive natural language driven data analyis platforms which would require sub-second latency.

7.2.1.4 Conclusion

Overall, open-source LLMs *can* achieve competitive NL2SQL performance as fine-tuning and in-context-learning prove to be effective and efficient mechanisms to achieve high accuracy values in constrained hardware environments. Nonetheless the measured performance yields the systems to face significant challenges in enterprise environments as an execution accuracy of 53.8% on BIRDwith a latency of 43.7s is unacceptable for end users. This indicates that while promising there is still a gap that remains to be closed.

7.2.2 Research Question 2: How do NL2SQL pipeline components interact, and which configurations optimize the accuracy-latency tradeoff?

7.2.2.1 Component Interaction Analysis

The pipeline components σ, ϕ, π, ρ and ν were shown to work best synergetically rather than additive with σ representing the primary performance driver of the pipeline nq. The ablation study perfomed during evaluation reveals this through a systematic evaluation of the three types of configurations: *Baseline*, *Zero-Shot* and *Full*. The evaluation results reveal the behaviour of the individual pipeline components in synergy, where *Zero-Shot* only provides a marginal improvement over *Baseline* (eg, +0.8pp on SPIDER) while nearly doubling the latency. All configurations with active ICL demonstrated significant improvement over their non-ICL counterparts; The worst results produced by ICL configurations compared to the best-performing non-ICL configurations resulted in an improvement through ICL by +1.7pp on SPIDER (dev), +2.1pp on SPIDER (test) and +14.4pp on BIRD. These results reveal that the other pipeline components bring only a minimal benefit (+0.8pp on SPIDER (dev), +0.5pp on SPIDER (test) and +1.6pp on BIRD) to the $\mathbb{EA}$ performance of a pipeline while significantly impacting (ie. approximately doubling) the $\mathbb{CL}$.

The underlying pattern of component interactions becomes even more apparent through examining NATURAL's failure modes. Without examples NATURAL (and thus OMNISQL) generate semantically inaccurate SQL candidates more often; usually incorrect column or table selection, misunderstanding of column semantics and value ranges or wrong join patterns result in syntactically correct queries which result in wrong result sets. For example the schema subsetting component is unable to work correctly if the initial query candidate is using the wrong set of tables alltogether, the majority voting algorithm fails to select the best candidate when all SQL candidates correspond to different result sets and self-refinement can only rarely repair semantically mislead

initial queries. In-context learning helps NATURAL to guide the foundational model by supplying a set of highly relevant, ranked examples which steers the model to put its attention on specific parts of the SQL schema. The schema-aware example selection algorithm works particularly well and was shown to be effective even on data that did not originate from the training splits of the corresponding benchmarks (1).

7.2.2.2 Component Contribution Ranking

The above insights allow for a ranking of components by their estimated impact based on the ablation evidence from Table 1 and Figure 7. The pipeline components rank by impact as follows:

1. **Query Projection (π)** – The query projection component itself remains the largest single contributor in performance and the foundation models used predominately drives the performance of the system. All other components are orchestrated around query projection. This was shown through the *Baseline* configuration which achieved 77.9% and 77.0% on SPIDER (dev and test) and 32.4% on BIRD. The substantial gap between the two benchmarks highlights how severly the model capabilities degrade with increasing task and schema complexity.

2. **Example Selection (σ)** – As discussed above, the example selection provides a non-neglible improvement in performance to the overall system while introducing relatively little candidate latency (eg, only +0.1-+2.3s on SPIDER). The impact that ICL showed on BIRDdemonstrates that example selection is not only helpful for LLMs generating SQL querys but a *necessary* measure for enabling them to handle complex query generation tasks on real-world data.

3. **Consensus Voting (ν)** – The consesus voting algorithm selects the most prevalent result set amogst multiple generation attempts and validates that queries execute without any errors prior to returning them to the user. This self-consensus mechanism ensures low syntactic error rates across all configurations (Figure 10). However the voting strategy only gains in overall performance impact with an increasing number of internal candidates. The evaluation was performed with 1-2 internal candidates due to computational complexity of increasing the pipeline iterations. This prevents definite conclusions on the impact.

4. **Self-Refinement (ρ)** – Self-refinement is effectively prompting the underlying model for repairing or improving its initially produced query candidate. This results in an overall marginal improvement of $\mathbb{EA}$, but comes at a significant computational cost. Adding self-refinement nearly doubles the $\mathbb{CL}$on average and provides diminishing returns. The results of the *Zero-Shot* configuration, especially on the BIRD benchmark, suggest that self-refinement is unable to correct significant semantic errors on its own without in-context learning.

5. **Schema-Subsetting (ϕ)** – Schema subsetting aims to reduce the noise for the model by filtering out unused table from the SQL schema before running query projection, thus improving token efficiency and accuracy. However, the exact impact of schema-subsetting is difficult to isolate due to the constrained ablation study that was performed. It is questionable whether schema-subsetting itself provides a statistically significant improvement in performance or even harms performance as the *Baseline* configuration underperformed the model baseline.

7.2.2.3 Accuracy-Latency Tradeoff

The accuracy-latency measurements which is visualized in Figure 9 and Table 1 reveal three different efficiency profiles across the systems that where benchmarked.

1. **Baseline** — The baseline profile includes both the *Baseline* configuration of NATURAL and the raw OMNISQL models. The baseline profile is generally yielding the lowest possible candidate latency, but also yields the lowest scores for both $\mathbb{EA}$ and $\mathbb{EM}$and is thus prioritising latency over accuracy.

2. **Zero-Shot** — The zero shot profile is denoted by a suboptimal accuracy latency tradeoff where there are marginal performance gain over the baseline results but this is accompanied by a strong average increase in latency (effectively doubling on average).

3. **Full** — Lastly the full profile is a balanced profile which weights accuracy over latency. Systems in this class include the *Syn*, *Train* and *Ground* configurations of NATURAL. These systems generally achieve best in class results but also a consistently and significantly higher candidate latency.

These three performance profiles highlight different tradeoffs that can be conciously made depending on the deployment scenario and the target usage pattern

7.2.2.4 Optimal Configurations for Different Deployment Scenarios

Given the tradeoffs a recommendation of profiles for certain deployment scenarios can be made. Generally the choice of the profile depends on the reliability requirements, the iteractivity of the use case and the complexity of the domain. Some examplary recommendations for common use cases can be made:

1. **Exploratory queries against simple databases** — For mostly interaction heavy systems like interactive query interfaces, data exploration and well-structured and smaller databases the **Baseline** configuration provides the overall best tradeoff given the lower latency which will significantly impact user experience. On simpler databases, the baseline configuration provides the best accuracy to latency ratio.
2. **Mixed complexity use cases** — For general purposed, mixed complexity use cases like internal analytical tools, asynchronous report generation and potentially complex databases the **Full** profile excels. If latency is generally less important than accuracy the latency problem can be eased by provisioning more powerful hardware, especially in enterprise or professional scenarios.
3. **Enterprise databases** — For large, complex databases with tens or even hundreds of tables, dirty data and potential external knowledge requirements (e.g, resemmbling BIRD) the **Full** profile is recommended. It provides meaningful assistance and paired with synthetic, historic and potentially domain-specialized examples it can achieve meaningful accuracy values in complex environments. Especially when manual querying would take significant amounts of time NATURAL can speed up the workflow.
4. **Research deployments** — For research deployments and user experience evaluations of NL2SQL databases the **Full** profile with potential ground truth samples is deemed optimal. NATURAL could serve for pure, theoretical evaluation of potential NL2SQL systems and their upper bound performance. Such a research deployment could reference past conversations and historical examples to form a self-learning system. This is configuration is not sensible for real world deployments as it would require a ground truth entry for all possibly asked questions.

Concluding, for highly interactive systems like chat applications and interactive learning tools, the **Baseline** profile will yield better latency characteristics while maintain a significant portion of accuracy, especially on simple tasks. For more complex usecases where accuracy is generally more critical than latency (e.g, asynchronous report generation), the **Full** profile is representing a more dedsirable tradeoff, which renders the **Zero-Shot** profile mostly undesirable.

7.2.2.5 Implications for NL2SQL Pipeline Design

General implications for future NL2SQL pipeline and system design can be derived from this work. First, the example selection and self-learning capabilities of any NL2SQL system that is based on LLMs will notably drive up it's accuracy. Focusing on example quality, example selection and example presentation will result in a good return of invest. Second, detailed component ablations will help to evaluate the concrete contribution of individual components, making iterations simpler and faster. Components should be evaluated against realistic benchmarks (e.g, BIRDor SPIDER2.0) to measure their real-world effects. Lastly, it is desirable to enable the system to be fine-tunable to an specific deployment scenario. Reference data used, fine-tuned models or specialized example generation can help to improve the accuracy scores on a concrete deployment case beyond the theoretical benchmark results that are cross-domain.

7.2.2.6 Conclusion

NATURALis a portable NL2SQL system which synergetically achieves largely competitive performance for a range of NL2SQL tasks. The **Full** profile of NATURAL was delivers significant accuracy gains over the **Baseline** profile on complex use cases. Nonetheless NATURAL suffers from poor latency characteristics over compared to the foundation models making it not always the better choice.

Therefore, the optimal NL2SQL system depends on the use case at hand, for enterprise level complexity using the advanced pipeline components can be beneficial, especially if continously learning systems are desirable. For exploratory and low latency systems, such as chatbots, NATURALprovides only marginal benefit over the underlying models.

A Appendix

A.1 Prompts

A.1.1 Natural Inference Prompt

```
EXAMPLE OF THE PROMPT
```

A.1.2 Natural Refinement Prompt

```
EXAMPLE OF THE PROMPT
```

A.1.3 OmniSQL Inference Prompt

```
EXAMPLE OF THE PROMPT
```

A.2 Code

A.2.1 Vector Database API

```
let question = "How many ..";

// Initialize vector database for similarity search
let vector = Vector::infer();
let model = EmbeddingModel::from_env().expect("failed to load embedding model");

// Query distance indices
let index = DistanceIndexRow::get("test_db", &vector)?;

// Embed question and perform similarity search
let selector = Selector::new(&question, None, &model, &vector)?;
let samples = selection::<4>(
    &Selector::new(
        question.as_ref(),
        None,
        &model,
        &vector,
    )?,
    &index.into(),
    &vector,
)?;
```

A.2.1.1 Vector Database Schema

```
CREATE TABLE IF NOT EXISTS databases (
  id               text not null,
  name             text not null,
  graph            text not null
);

CREATE TABLE IF NOT EXISTS distance_indices (
  db_id            text not null,
  distances        text not null
);

CREATE VIRTUAL TABLE IF NOT EXISTS samples USING vec0 (
  id                 text not null,
  db_id              text not null,
```

```
  question           text not null,
  question_embedding float[2048],

  sql                text not null,
  sql_embedding      float[2048]
);
```

A.2.1.2 Benchmark Traits

```
pub trait Benchmark {
    type Error: From<eyre::Report> + Debug;
    type Database: BenchmarkDatabase<Error = Self::Error>;

    fn name() -> &'static str;

    fn databases(&self) -> impl Iterator<Item = &Self::Database>;
    fn database(&self, id: impl AsRef<str>) -> Option<&Self::Database>;

    fn tests(&self) -> impl Iterator<
      Item = &<Self::Database as BenchmarkDatabase>::Test
    >;

    fn test(&self, id: impl AsRef<str>) -> Option<
      &<Self::Database as BenchmarkDatabase>::Test
    >;
}

pub trait BenchmarkDatabase {
    type Error: From<eyre::Report> + Debug;
    type Test: BenchmarkTest;

    fn id(&self) -> &str;
    fn connect(&self) -> Result<Database, Self::Error>;
    fn tests(&self) -> impl Iterator<Item = &Self::Test>;
    fn test(&self, id: impl AsRef<str>) -> Option<&Self::Test>;
}

pub trait BenchmarkTest {
    fn id(&self) -> &str;
    fn db_id(&self) -> &str;

    fn name(&self) -> &str;
    fn question(&self) -> &str;
    fn query(&self) -> &str;
}
```

A.3 Benchmark Results

Detailed measurements and visualizations for the evaluation section.

A.3.1 Spider

This section provides detailed breakdowns of the Spider benchmark results for both development and test splits, showing execution accuracy and exact match metrics across all system configurations.

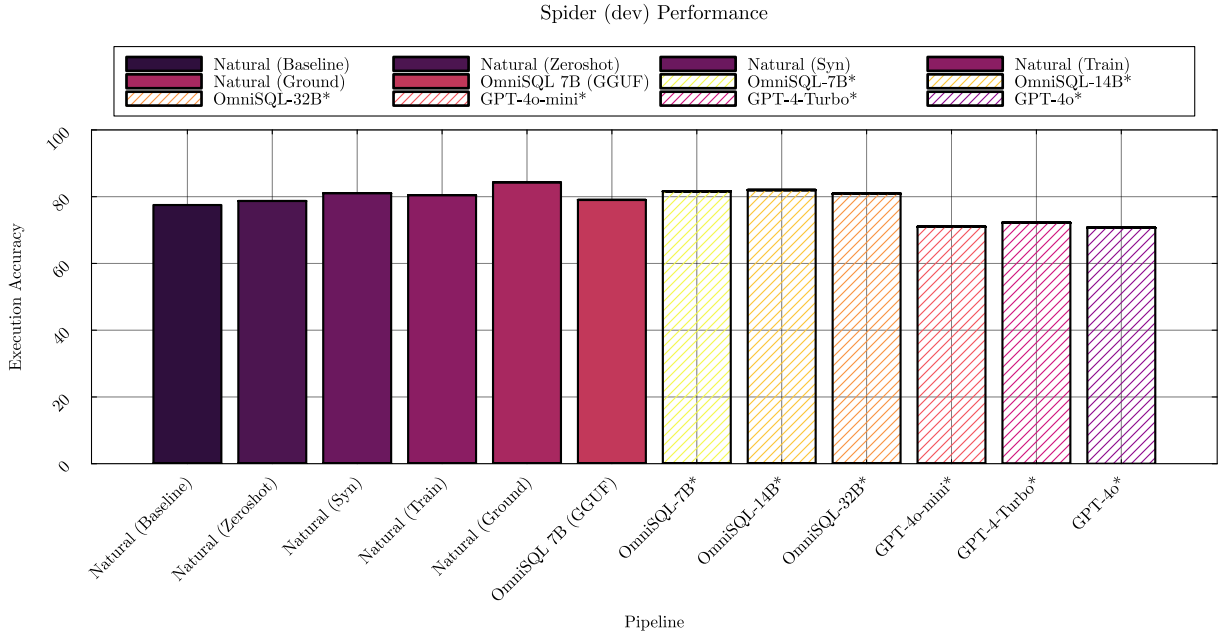

Figure 12: Execution accuracy on SPIDER (dev). The *Syn* configuration achieves the highest execution accuracy among all tested systems at 81.0%, outperforming OmniSQL-7B-gguf by 2.0 percentage points.

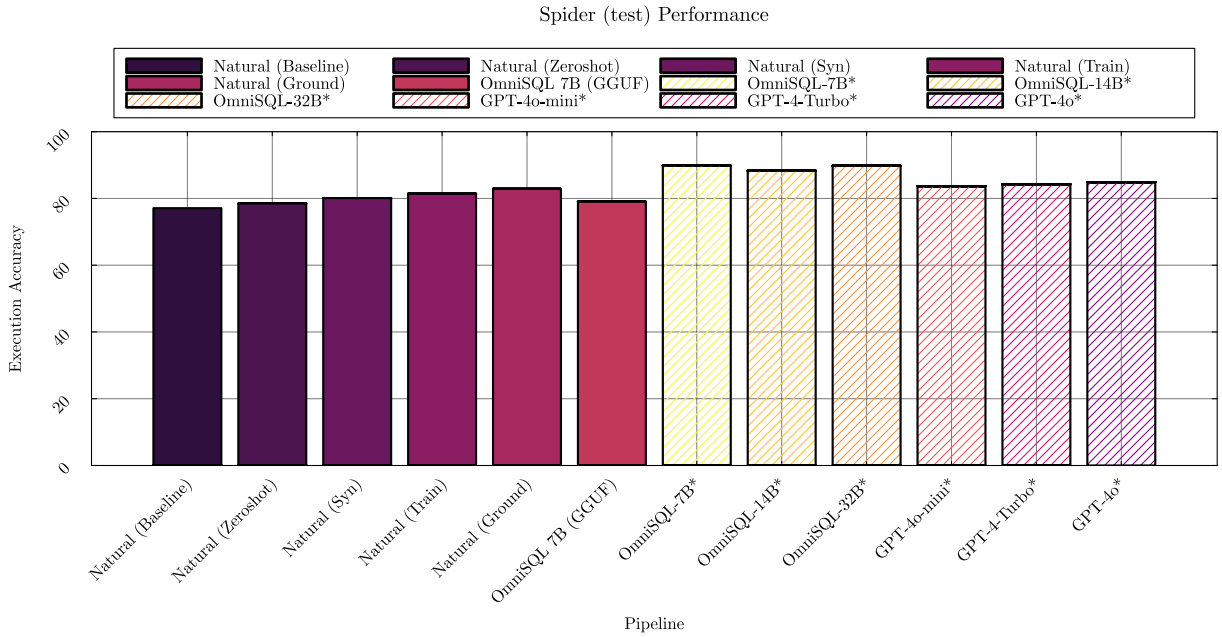

Figure 13: Execution accuracy on SPIDER (test). The *Train* configuration demonstrates the strongest performance among tested configurations, achieving 81.4% execution accuracy.

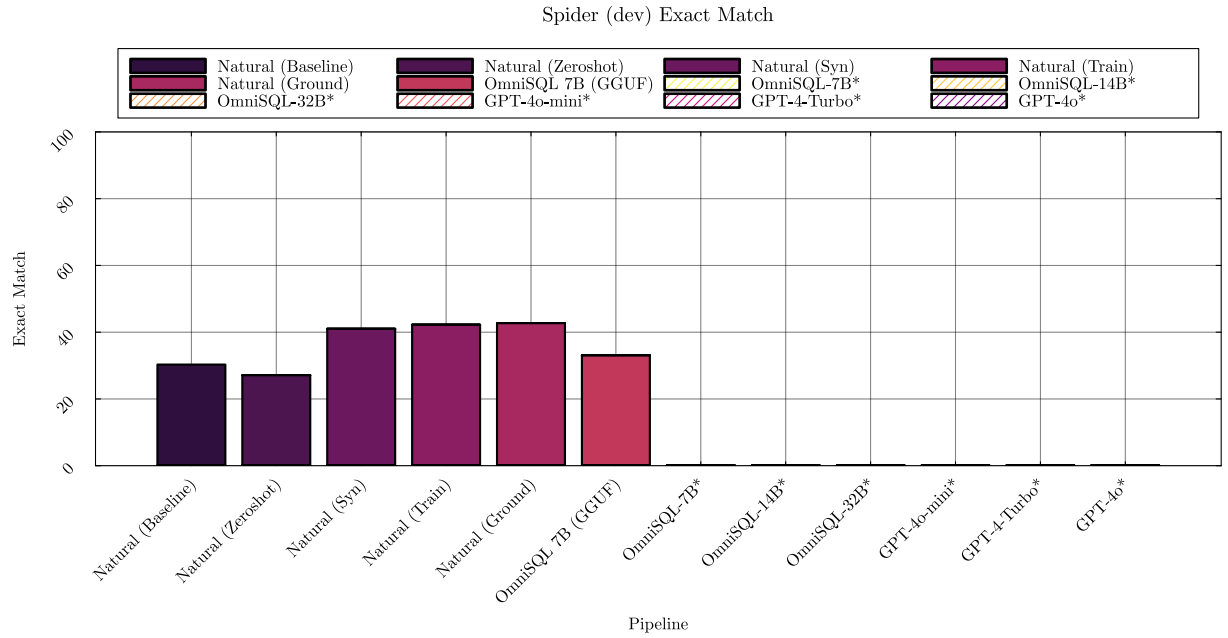

Figure 14: Exact match performance on Spider (dev). ICL configurations demonstrate substantial improvements over baseline, with *Ground* achieving 42.7% exact match compared to 30.3% for *Baseline*.

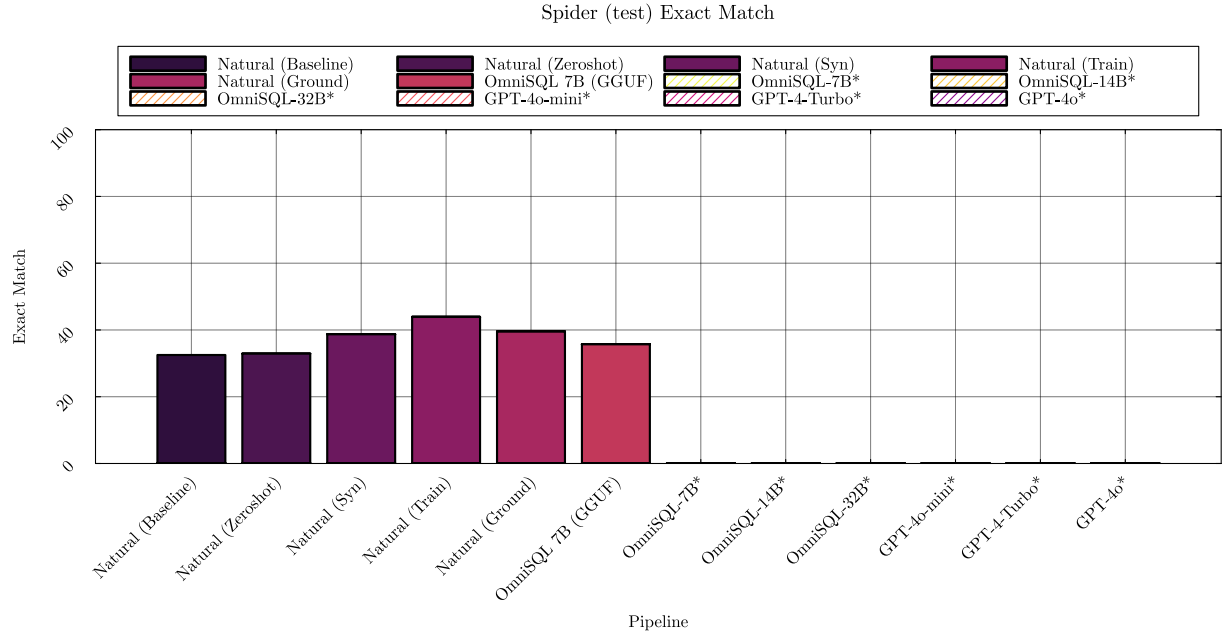

Figure 15: Exact match performance on Spider (test). The *Train* configuration achieves 43.9% exact match, representing a 21.4 percentage point improvement over *Baseline*.

A.3.2 Bird

This section provides detailed breakdowns of the BIRD benchmark results its development split, showing execution accuracy and exact match metrics across all system configurations.