Database Education in the Era of GenAI

INDE lab

University of Amsterdam
Informatics Institute

Daphne Miedema

Positioning of my work

Positioning of my work

Identifying SQL Misconceptions of Novices: Findings from a Think-Aloud Study

Daphne Miedema
d.e.miedema@tue.nl
Artificial Intelligence and Data Engineering (AIDE) Lab
Eindhoven, the Netherlands
Eindhoven University of Technology
Eindhoven, the Netherlands

Efthimia Aivaloglou
e.aivaloglou@liacs.leidenuniv.nl
Leiden Institute of Advanced Computer Science
Leiden, The Netherlands
Open Universiteit
Heerlen, The Netherlands

George Fletcher
g.h.l.fletcher@tue.nl
Artificial Intelligence and Data Engineering (AIDE) Lab
Eindhoven, the Netherlands
Eindhoven University of Technology
Eindhoven, the Netherlands

ABSTRACT

SQL is the most commonly taught database query language. While previous research has investigated the errors made by novices during SQL query formulation, the underlying causes for these errors have remained unexplored. Understanding the basic misconceptions held by novices which lead to these errors would help improve how we teach query languages to our students. In this paper we aim to identify the misconceptions that might be the causes of documented SQL errors that novices make. To this end, we conducted a qualitative think-aloud study to gather information on the thinking process of university students while solving query formulation problems. With the queries in hand, we analyzed the underlying causes for the errors made by our participants. In this paper we present the identified SQL misconceptions organized into four top-level categories: misconceptions based in previous course knowledge, generalization-based misconceptions, language-based misconceptions, and misconceptions due to an incomplete or incorrect mental model. A deep exploration of misconceptions can uncover gaps in instruction. By drawing attention to these, we aim to improve SQL education.

1 INTRODUCTION

Databases and the Structured Query Language (SQL) are core topics in Software Engineering and Computer Science curricula in higher education [33]. While the syntax of SQL is relatively simple in comparison to most programming languages, SQL is not a trivial language to learn. Several works have reported on common errors in the SQL queries written by students and novices, identifying various categories such as syntax errors, logical errors, semantic errors and complications [3, 8, 21, 31, 32, 34]. In the area of research surrounding SQL education, the main focus has been on the errors made, without expanding on the underlying causes or the misconceptions that novices may hold. The identification of misconceptions is an important first step towards devising instructional approaches that address students' difficulties. In the area of programming education, on the other hand, misconceptions have been well studied. Even though SQL is a query language, not a programming language, parallels between the two can be drawn.

Early works on programming misconceptions identified them as conceptual bugs in how novices program and understand programs [20], difficulties of learning to program, and errors based on the misapplication of analogies [7]. Since then, a large body of work has

Positioning of my work

“There is no ambiguity on what to return”: Investigating the Prevalence of SQL Misconceptions

Daphne Miedema
d.e.miedema@tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

Michael Liut
michael.liut@utoronto.ca
University of Toronto Missisauga
Missisauga, Canada

George Fletcher
g.h.l.fletcher@tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

Efthimia Aivaloglou
e.aivaloglou@tudelft.nl
Delft University of Technology
Delft, The Netherlands

ABSTRACT

In recent years, database education has been receiving more attention, with research in various directions such as the development of tools for education, the analysis of students’ homework, and the exploration of misconceptions. Misconceptions are mistakes in student reasoning that lead to errors during problem-solving. Recent work has documented misconceptions and errors in SQL. In this study we test the prevalence of several of these misconceptions through a multiple-choice questionnaire, to see if they hold on a larger, more diverse, student population. We found that all misconceptions are held to some extent, with prevalence scores ranging from one to fifty-two percent of the student population. Additionally, we have uncovered previously unidentified areas of struggle, allowing us to identify new misconceptions.

CCS CONCEPTS

• **Information systems** → **Structured Query Language**; • **Social and professional topics** → **Computing education**.

tools have been developed for supporting students in query writing [16, 28, 32, 33] and repair [26, 31], and visualisations have been proposed for assisting in query understanding [14, 23, 28]. Several existing works have concluded that SQL, simple as it may appear, is challenging for novices.

Recent research has identified misconceptions as underlying causes of student challenges and errors. This line of work was initiated by Taipalus [49], who also discussed the potential causes behind persistent SQL errors [50], and was continued by Miedema et al. [27], who collected and analyzed qualitative data on the thought process of 21 students while they were working on query formulation problems. In their work, they identified fourteen misconceptions in four categories: misconceptions based on previous course knowledge, generalization-based misconceptions, misconceptions based on language, and misconceptions due to an incomplete or incorrect mental model [27].

This paper builds on that work. It is inspired by the importance of research on understanding the problems novices face with SQL, and the design of interventions to support them. Currently, even though

Early works on programming misconceptions identified them as conceptual bugs in how novices program and understand programs [20], difficulties of learning to program, and errors based on the misapplication of analogies [7]. Since then, a large body of work has

Positioning of my work

Students' Perceptions on Engaging Database Domains and Structures

Daphne Miedema
d.e.miedema@tue.nl
Eindhoven University of Technology
Eindhoven, the Netherlands

Toni Taipalus
toni.taipalus@jyu.fi
University of Jyväskylä
Jyväskylä, Finland

Efthimia Aivaloglou
e.aivaloglou@tudelft.nl
Open University
Heerlen, the Netherlands
Delft University of Technology
Delft, the Netherlands

ABSTRACT

Several educational studies have argued for the contextualization of assignments, i.e., for providing a context or a story instead of an abstract or symbolic problem statement. Such contextualization may have beneficial effects such as higher student engagement and lower dropout rates. In the domain of database education, textbooks and educators typically provide an example database for context. These are then used to introduce key concepts related to database design, and to illustrate querying. However, it remains unstudied what kinds of database contexts are engaging for novices. In this paper, we study which aspects of database domain and complexity students find engaging through student reflections on a database creation assignment. We identify six factors regarding engaging domains, and five factors for engaging complexity. The main factor for domain-related engagement was *Personal interest*, the main factor for complexity engagement was *Matching information requirements*. Our findings can help database educators and book authors to design engaging exercise databases targeted for novices.

CCS CONCEPTS

• **Applied computing** → **Education**; • **Information systems** → **Relational database model**.

without data. As such, data management techniques are also an essential part of computer science education.

Data management is typically taught by means of toy examples [23], with domains such as store ordering systems, movie rentals and company employees [21]. These examples are accessible: most students have some idea of what data could be involved in these domains, which they can use as a scaffold to remember the database schema. As such, the domain of the data is the inherent context of the database, but determining what makes a good database context not been studied in-depth before.

We can deepen students interest in course material by ensuring that the projects they work on are authentic, and by giving them a choice in the what and how of the project [2]. For example, teaching CS1 in the students major's context has led to increased student success rates [22]. Increased interest in a project may also lead the students to spend more time on an assignment [6]. Furthermore, increased interest makes students become more active learners, which in turn increases motivation and learning [11]. However, it has been shown that more complex databases have the possible downside of making learning SQL more difficult [21].

In this paper, we examine factors that students think make databases engaging, to answer the question: *What kind of databases do novices deem engaging to study database design, implementation*

"There i

Eindh
E

Eindh
E

ABSTRACT

In recent years, datab
tion, with research ir
of tools for educatio
the exploration of m
in student reasoning
Recent work has doc
In this study we test
tions through a mult
on a larger, more div
misconceptions are
ranging from one to
Additionally, we hav
struggle, allowing us

CCS CONCEPTS

• **Information systems** → **Structured Query Language**; • **Social and professional topics** → **Computing education**.

incorrect mental model [27].

This paper builds on that work. It is inspired by the importance of research on understanding the problems novices face with SQL, and the design of interventions to support them. Currently, even though

Early works on programming misconceptions identified them as conceptual bugs in how novices program and understand programs [20], difficulties of learning to program, and errors based on the misapplication of analogies [7]. Since then, a large body of work has

Identi

Daj
d.e.
Artificial
Engi
Eindho
Eindhoven
Eindho

ABSTRACT

SQL is the most
previous research
ing SQL query fo
have remained u
tions held by nov
how we teach qu
aim to identify
documented SQ
ducted a qualita
the thinking pro
formulation prob
underlying cause
paper we present
four top-level ca
knowledge, gene
misconceptions,
correct mental model. A deep exploration of misconceptions can uncover gaps in instruction. By drawing attention to these, we aim to improve SQL education.

Positioni

Staring at Tables: Exploring Conceptual Data Modeling as a Rich Collaborative Activity

Laura Koesten
MBZUAI & University of Vienna & AIT
Abu Dhabi, UAE / Austria
laura.koesten@mbzuai.ac.ae

Daphne Miedema
University of Amsterdam
Amsterdam, the Netherlands
d.e.miedema@uva.nl

Hsiang-Yun Wu
University of Applied Sciences St. Pölten & TU Wien
St. Pölten / Vienna, Austria
hsiang-yun.wu@ustp.at

Mathias Funk
Eindhoven University of Technology
Eindhoven, the Netherlands
m.funk@tue.nl

Figure 1: Three examples of conceptual models resulting from our study (selected for diversity); Left: Style of an ER diagram (D2) ; Center: Informal style (D8) ; Right: Style of a star schema (D11)

Abstract

Conceptual data modeling is a central activity in data work, yet how such models are created remains understudied. While data

CCS Concepts

• **Human-centered computing** → *Collaborative content creation*; *User studies*; • **Information systems** → *Database design and mod-*

has been shown that more complex databases have the possible downside of making learning SQL more difficult [21].

In this paper, we examine factors that students think make databases engaging, to answer the question: *What kind of databases do novices deem engaging to study database design, implementation*

incorrect mental model [27].

This paper builds on that work. It is inspired by the importance of research on understanding the problems novices face with SQL, and the design of interventions to support them. Currently, even though

ABSTRACT

Several educational stu
of assignments, i.e., fo
an abstract or symbolic
may have beneficial eff
lower dropout rates. In
and educators typicall
These are then used to
design, and to illustrat
what kinds of databas
paper, we study which
students find engaging
creation assignment.
domains, and five fact
tor for domain-related
factor for complexity
quirements. Our findi
authors to design engaging exercise databases targeted for novices.

CCS CONCEPTS

• **Applied computing** → **Education**; • **Information systems** → **Relational database model**.

ABSTRACT

In recent years, datab
tion, with research in
of tools for educatio
the exploration of m
in student reasoning
Recent work has doc
In this study we test
tions through a mult
on a larger, more div
misconceptions are
ranging from one to
Additionally, we hav
struggle, allowing us

CCS CONCEPTS

• **Information systems** → **Structured Query Language**; • **Social and professional topics** → **Computing education**.

Early works on programming misconceptions identified them as conceptual bugs in how novices program and understand programs [20], difficulties of learning to program, and errors based on the misapplication of analogies [7]. Since then, a large body of work has

ABSTRACT

SQL is the most
previous researc
ing SQL query fo
have remained
tions held by nov
how we teach qu
aim to identify
documented SQ
ducted a qualita
the thinking pro
formulation prob
underlying cause
paper we present
four top-level ca
knowledge, gene
misconceptions,
correct mental model. A deep exploration of misconceptions can uncover gaps in instruction. By drawing attention to these, we aim to improve SQL education.

Positioni

Identi

Da
d.e.
Artificial
Engi
Eindho
Eindhoven
Eindho

ABSTRACT

SQL is the most
previous researc
ing SQL query fo
have remained u
tions held by nov
how we teach qu
aim to identify
documented SQ
ducted a qualita
the thinking pro
formulation prob
underlying cause
paper we present
four top-level ca
knowledge, gene
misconceptions,
correct mental model. A deep exploration of misconceptions can uncover gaps in instruction. By drawing attention to these, we aim to improve SQL education.

"There i

Eindh
E
Eindh
E

ABSTRACT

In recent years, datab
tion, with research ir
of tools for educatio
the exploration of m
in student reasoning
Recent work has doc
In this study we test
tions through a mult
on a larger, more div
misconceptions are
ranging from one to
Additionally, we hav
struggle, allowing us

CCS CONCEPTS

• **Information systems** → **Structured Query Language**; • **Social and professional topics** → **Computing education**.

Early works on programming misconceptions identified them as conceptual bugs in how novices program and understand programs [20], difficulties of learning to program, and errors based on the misapplication of analogies [7]. Since then, a large body of work has

Staring at

Daphne
d.e.mied
Eindhoven Unive
Eindhoven,

ABSTRACT

Several educational st
of assignments, i.e., fo
an abstract or symboli
may have beneficial eff
lower dropout rates. In
and educators typicall
These are then used to
design, and to illustrat
what kinds of databas
paper, we study which
students find engaging
creation assignment.
domains, and five fact
tor for domain-related
factor for complexity
quirements. Our findi
authors to design engaging exercise databases targeted for novices.

CCS CONCEPTS

• **Applied computing** → **Education**; • **Information systems** → **Relational database model**.

MBZUA
A
la

University of A
St

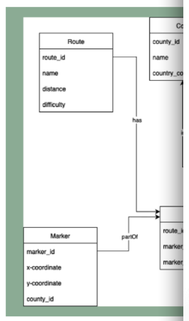

Figure 1: Three exa (D2) ; Center: Informal style (D8) ; Right: Style of a star schema (D11)

Abstract

Conceptual data modeling is a central activity in data work, yet how such models are created remains understudied. While data

has been shown that more complex databases have the possible downside of making learning SQL more difficult [21].

In this paper, we examine factors that students think make databases engaging, to answer the question: *What kind of databases do novices deem engaging to study database design, implementation*

incorrect mental model [27].

This paper builds on that work. It is inspired by the importance of research on understanding the problems novices face with SQL, and the design of interventions to support them. Currently, even though

CCS Concepts

• **Human-centered computing** → *Collaborative content creation*; *User studies*; • **Information systems** → *Database design and mod-*

A Feasibility Study on Automated SQL Exercise Generation with ChatGPT-3.5

Willem Aerts
Eindhoven University of Technology
Eindhoven, The Netherlands
willem.aerts@onsbrabantnet.nl

George Fletcher
Eindhoven University of Technology
Eindhoven, The Netherlands
g.h.l.fletcher@tue.nl

Daphne Miedema
Eindhoven University of Technology
Eindhoven, The Netherlands
d.e.miedema@tue.nl

ABSTRACT

SQL is the standard for database query languages and is taught in most introductory database courses. Query languages are illustrated and tested through toy examples: small, accessible, instances of databases. These are not always engaging, but coming up with new examples and questions is time-consuming. Existing research in Computer Science Education has shown that Large Language Models (LLMs) can generate coding exercises. However, this has not been demonstrated for SQL yet but could save teachers much time. In this paper, we study whether it is feasible to have ChatGPT-3.5 generate database schemas and associated SQL questions for teachers through a two-part study.

Through a survey of educators, we found that creating a story and database schema for the SQL part is more time-consuming than the questions themselves. In our prompt engineering study, we identified prompts that were successful at creating database schemas, mock data, and exercises. However, although ChatGPT could help reduce the time required to create exams, some participants indicated that they are skeptical about using LLMs.

CCS CONCEPTS

• **Computing methodologies** → **Generative and developmental approaches**; • **Information systems** → **Structured Query Language**; • **Social and professional topics** → **Information systems education**.

Structured Query Language, or SQL. It is an important topic to teach students well, as it is the number one requested language for computer science jobs [5], and is applicable in many other fields too such as business analytics [9].

To assess whether students understand SQL well enough, teachers create homeworks, assignments, and final exams. In the case of SQL, this involves the creation of a database schema and mock data before being able to create actual questions. This design is a time-consuming task, especially if teachers aim to significantly vary their databases between iterations of the course.

Given previous work on the generation of exercises for programming [21, 29], we wondered if a similar task would be possible for data systems education. In this paper, we investigate which SQL concepts data systems educators test, and whether it is possible for LLMs to help teachers create questions on these. We pose the following research questions:

RQ1 Do data systems educators see the advantage of using LLMs to generate exercises?

RQ2 Can ChatGPT-3.5 generate appropriate sets of schema, data and SQL questions for teachers to use?

2 RELATED WORK

2.1 Data Systems Education

Data Systems Education research has become increasingly popular over the past years. One core topic is the identification and organi-

Data Systems Education

There have always been challenges

Some known challenges in education

Coming from imperative languages where problem decomposition is straightforward, the declarative nature of SQL can be uncomfortable.

For most students, databases are a far cry from their favourite courses, leaving them struggling to engage.

The non-determinism associated with query execution plans can be confusing.

The fact that SQL uses three-valued logic is unknown and/or misunderstood.

SQL errors are common: Semantic, Syntactic, Logic

Errors are caused by misconceptions: misunderstandings of how the underlying technology (such as the use of SQL clauses) works. Misconceptions are difficult to mitigate.

Error messages are often incomplete and require reading manuals to understand.

DMBS will only report the first error in the query, and thus make it harder to debug the query as a whole.

GenAI for education is here

Students can offload their education

GenAI can pass assessments…

- MCQs on programming
  - GPT-3 answered 37.5%
  - GPT-3.5 answered 64.3%
  - GPT-4 answered 84.1%
- GPT-4 scores 71% on open coding questions.
- GPT-4 can pass an intermediate programming course.

Jaromir Savelka, Arav Agarwal, Marshall An, Chris Bogart, and Majd Sakr. 2023. Thrilled by Your Progress! Large Language Models (GPT-4) No Longer Struggle to Pass Assessments in Higher Education Programming Courses. ICER 2023. https://doi.org/10.1145/3568813.3600142

- Students create a database for a bookstore by completing the following learning activities:
  - (a) Write the SQL schema for creating the database from the given logical model.
  - (b) Write SQL statements for populating the schema with data.
  - (c) Build the bookstore database in MySQL.
  - (d) Write SQL queries on the bookstore database for retrieving specific information.
- Students were permitted to use ChatGPT to assist.

P. Lauren and P. Watta, "Work-in-Progress: Integrating Generative AI with Evidence-based Learning Strategies in Computer Science and Engineering Education," *2023 IEEE Frontiers in Education Conference (FIE).* doi: 10.1109/FIE58773.2023.10342970.

GenAI can pass assessments…

- MCQs on programming
  - GPT-3 answered 37.5%
  - GPT-3.5 answered 64.3%
  - GPT-4 answered 84.1%
- GPT-4 scores 71% on open coding questions.
- GPT-4 can pass an intermediate programming course.

- Students create a database for a bookstore by c
- S ist.

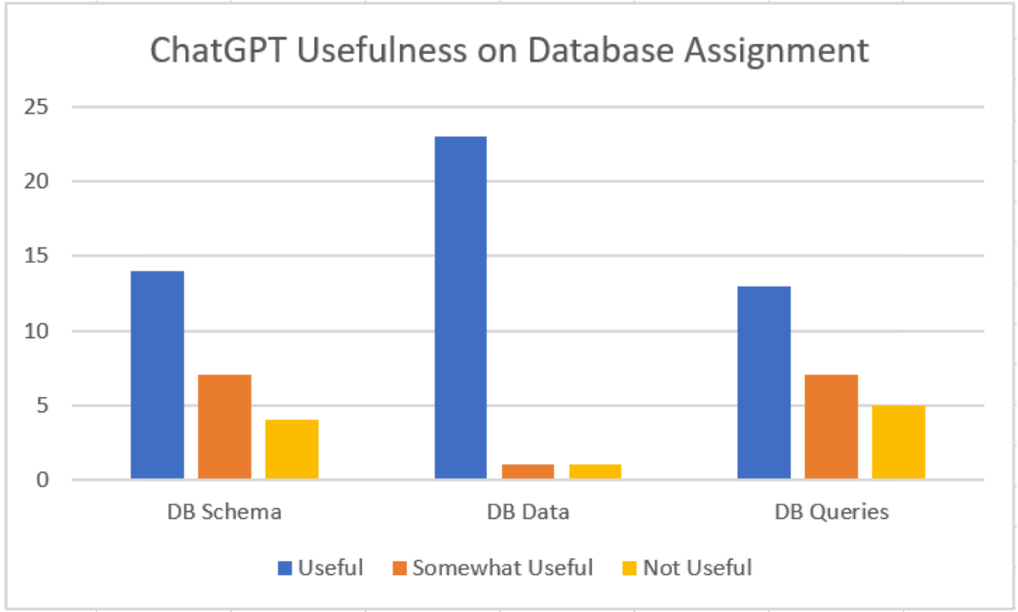

Fig. 1. ChatGPT Usefulness on MCS 3543 Assignment.

Jaromir Savelka, Arav Agarwal, Marshall An, Chris Bogart, and Majd Sakr. 2023. Thrilled by Your Progress! Large Language Models (GPT-4) No Longer Struggle to Pass Assessments in Higher Education Programming Courses. ICER 2023. https://doi.org/10.1145/3568813.3600142

P. Lauren and P. Watta, "Work-in-Progress: Integrating Generative AI with Evidence-based Learning Strategies in Computer Science and Engineering Education," *2023 IEEE Frontiers in Education Conference (FIE).* doi: 10.1109/FIE58773.2023.10342970.

… and it will give away the answers

- In a study where students programmed 45 Python scripts across 10 sessions, in 49% of cases AI-generated code was submitted by students without any modification.

Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J. Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the Effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. CHI 2023. https://doi.org/10.1145/3544548.3580919

- In response to help-requests by students, 99% of GPT-3.5 responses contained source code, even when directly asked not to.

Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. 2023. Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. ICER 2023. https://doi.org/10.1145/3568813.3600139

… and it will give away the answers

- In a study where students programmed 45 Python scripts across 10 sessions, in 49% of cases AI-generated code was submitted by students without any modification.

Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J. Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the Effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. CHI 2023. https://doi.org/10.1145/3544548.3580919

- In response to help-requests by students, 99% of GPT-3.5 responses contained source code, even when directly asked not to.

Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. 2023. Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. ICER 2023. https://doi.org/10.1145/3568813.3600139

- Better guardrails are required to avoid sharing answers!

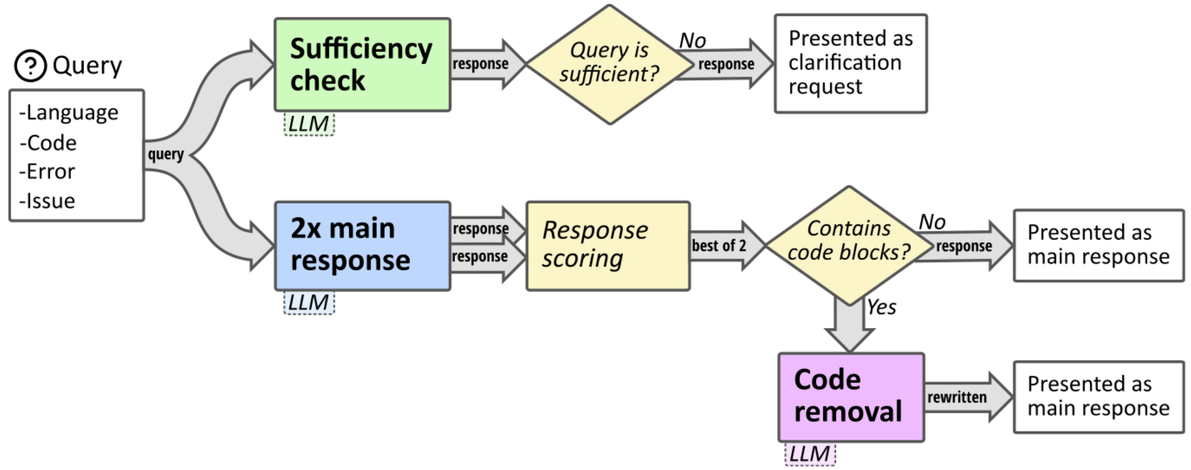

Figure 4: CodeHelp's response workflow. Steps using a large language model completion are tagged *LLM*.

Mark Liffiton, Brad E Sheese, Jaromir Savelka, and Paul Denny. 2024. CodeHelp: Using Large Language Models with Guardrails for Scalable Support in Programming Classes. Koli Calling 2023. https://doi.org/10.1145/3631802.3631830

UvA AI Chat

Welcome to UvA AI Chat - the generative AI application of University of Amsterdam that supports you in your studies and education. Please be aware that the generated information may not always be complete, correct or up-to-date, so remain critical and use your own judgment. Personal data is handled with care: only your account details are stored, and you can easily delete your conversations and documents. Please use the chat in accordance with the University of Amsterdam's guidelines.

Try one of these prompts:

- Create a study plan
- Generate study questions
- Explain a difficult concept
- Review my essay

Type your message here...

gpt-5.1

Upside: GenAI as tutor

Fig. 3. User has requested an exercise question, and MoodleBot is assisting with the solution.

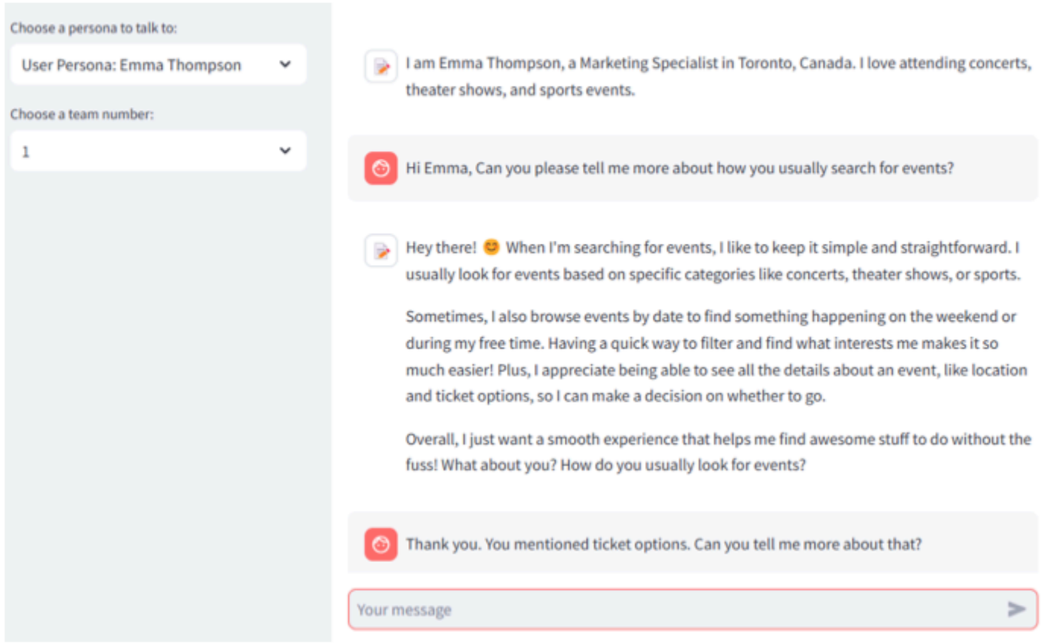

Figure 1: User interface of the Requirement Elicitation Tool

A. T. Neumann, Y. Yin, S. Sowe, S. Decker and M. Jarke, 2025. An LLM-Driven Chatbot in Higher Education for Databases and Information Systems. IEEE Transactions on Education. doi: 10.1109/TE.2024.3467912.

Ildar Akhmetov and Mirjana Prpa. 2025. Simulating Requirement Elicitation: Development and Evaluation of a Persona-Based Tool. SIGCSE 2025. https://doi.org/10.1145/3641555.3705250

Downside: messes with comprehension and metacognition

- Good students are able to use GenAI to accelerate their progress.
- Struggling students experience persisting metacognitive difficulties and cognitive dissonance.
- Difficulties can be compounded by GenAI (especially Location)
- Users report time saving as a benefit of GenAI, but this is not supported by research findings.

Table 2: Definitions of Old and New Metacognitive Difficulties

Name	Description
	Previous [52]
Forming	Forming the wrong conceptual model about the right problem.
Dislodging	Dislodging an incorrect conceptual model of the problem may not be solved.
Assumption	Forming the correct conceptual model for the wrong problem.
Location	Moving too quickly through one or more stages incorrectly leads to a false sense of accomplishment and poor conception of location in the problem-solving process.
Achievement	Unwillingness to abandon a wrong solution due to a false sense of being nearly done.
	New
Progression	Being conceptually behind in the course material but unaware of it due to a false sense of confidence
Interruption	An inability to concentrate on problem solving due to frequent interruptions and code suggestions.
Mislead	The tool leads the user down the wrong path.

James Prather, Brent N Reeves, Juho Leinonen, Stephen MacNeil, Arisoa S Randrianasolo, Brett A. Becker, Bailey Kimmel, Jared Wright, and Ben Briggs. 2024. The Widening Gap: The Benefits and Harms of Generative AI for Novice Programmers. ICER 2024. https://doi.org/10.1145/3632620.3671116

Why do students use GenAI?

ChatGPT and other LLMs

- More explanations
- Simpler explanations
- Supplement course material
- Ability to ask questions
- Accessibility (f.e. non-native speakers)
- Brainstorming
- Busy schedules

No LLM

- Course materials are sufficient
- LLMs provide incorrect answers
- LLMs hallucinate
- Practice is key to success

Valeria Ramirez Osorio, Angela Zavaleta Bernuy, Bogdan Simion, and Michael Liut. 2025. Understanding the Impact of Using Generative AI Tools in a Database Course. SIGCSE 2025. https://doi.org/10.1145/3641554.3701785

Suggestions for educational practice

Old challenges remain, new challenges arise

Engagement

Miedema, D., Taipalus, T., & Aivaloglou, E. (2023). Students' Perceptions on Engaging Database Domains and Structures. SIGCSE 2023. https://doi.org/10.1145/3545945.3569727

Taipalus, T., Miedema, D., & Aivaloglou, E. (2023). Engaging Databases for Data Systems Education. ITiCSE 2023. https://doi.org/10.1145/3587102.3588804

Engagement

Miedema, D., Taipalus, T., & Aivaloglou, E. (2023). Students' Perceptions on Engaging Database Domains and Structures. SIGCSE 2023. https://doi.org/10.1145/3545945.3569727

Taipalus, T., Miedema, D., & Aivaloglou, E. (2023). Engaging Databases for Data Systems Education. ITiCSE 2023. https://doi.org/10.1145/3587102.3588804

- Familiarity
- Connection to practice
- Learning opportunity
- Versatility
- Social and ethical issues
- Practical constraints
- Simplicity

Engagement

Miedema, D., Taipalus, T., & Aivaloglou, E. (2023). Students' Perceptions on Engaging Database Domains and Structures. SIGCSE 2023. https://doi.org/10.1145/3545945.3569727

Taipalus, T., Miedema, D., & Aivaloglou, E. (2023). Engaging Databases for Data Systems Education. ITiCSE 2023. https://doi.org/10.1145/3587102.3588804

Table 1: Database domains and purposes categorized into four themes; note that some domains pertain to more than one category, e.g., there were five databases for digital music platforms, and three of these five databases also contained data structures for social interaction

Database type (# of occurrences)	Database domains (# of occurrences, if more than one)
Support for a physical service (24)	dog contest and health (3); books (2); car sales (2); hotel reservation system (2); university course enrollment (2); bank; board game details and interrelationships; business trip invoicing; car rental service; car wash; employee management; gym memberships; hospital access control; library; multidisciplinary primary school courses; pet health; statistics on board game matches; vaccinations
Delivery of physical or digital goods (21)	online shop (7); digital music platform (5); digital video game distribution platform (4); mobile application store; food delivery platform; marketplace for internet domains; online multiplayer game; operating system update service
Information propagation or collection (14)	statistics on soccer matches (3); dog contest and health (3); academic publications (2); concerts; digital game speedruns; digital mobile gamers; music and movie streaming; pet health; trekking locations
Social interaction (8)	digital video game distribution platform (4); digital music platform (3); car sales

- Familiarity
- Connection to practice
- Learning opportunity
- Versatility
- Social and ethical issues
- Practical constraints
- Simplicity

Engagement

Miedema, D., Taipalus, T., & Aivaloglou, E. (2023). Students' Perceptions on Engaging Database Domains and Structures. SIGCSE 2023. https://doi.org/10.1145/3545945.3569727

Taipalus, T., Miedema, D., & Aivaloglou, E. (2023). Engaging Databases for Data Systems Education. ITiCSE 2023. https://doi.org/10.1145/3587102.3588804

Table 1: Database domains and purposes categorized into four themes; note that some domains pertain to more than one category, e.g., there were five databases for digital music platforms, and three of these five databases also contained data structures for social interaction

Database type (# of occurrences)	Database domains (# of occurrences, if more than one)
Support for a physical service (24)	dog contest and health (3); books (2); car sales (2); hotel reservation system (2); university course enrollment (2); bank; board game details and interrelationships; business trip invoicing; car rental service; car wash; employee management; gym memberships; hospital access control; library; multidisciplinary primary school courses; pet health; statistics on board game matches; vaccinations
Delivery of physical or digital goods (21)	online shop (7); digital music platform (5); digital video game distribution platform (4); mobile application store; food delivery platform; marketplace for internet domains; online multiplayer game; operating system update service
Information propagation or collection (14)	statistics on soccer matches (3); dog contest and health (3); academic publications (2); concerts; digital game speedruns; digital mobile gamers; music and movie streaming; pet health; trekking locations
Social interaction (8)	digital video game distribution platform (4); digital music platform (3); car sales

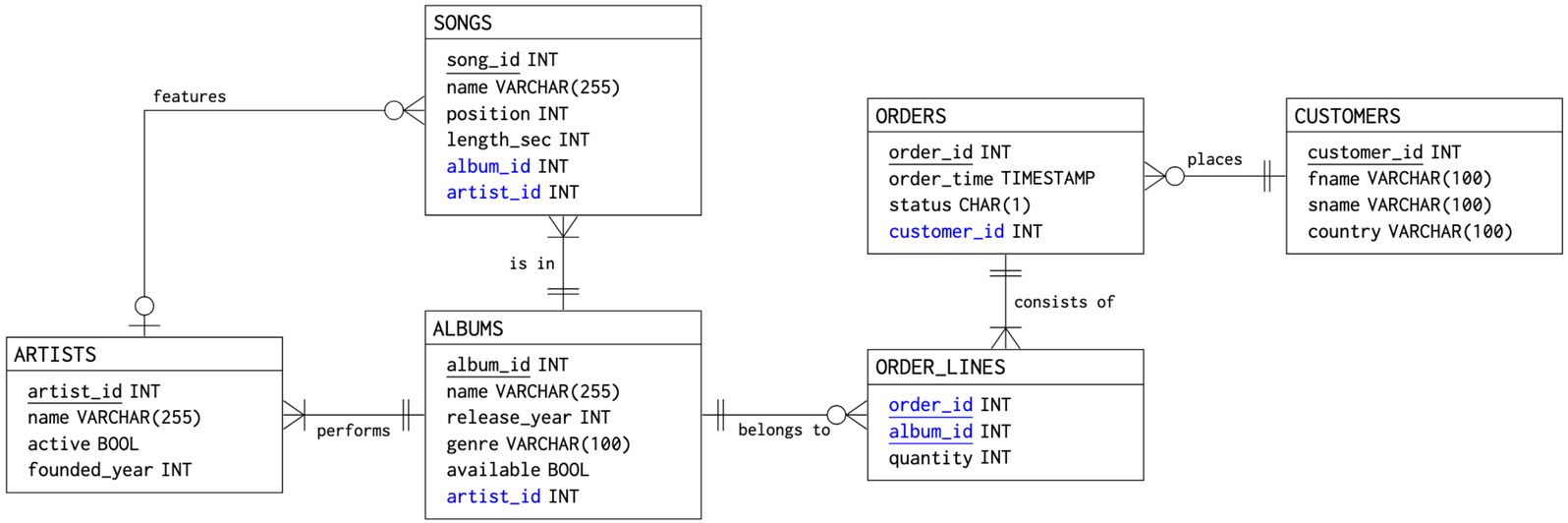

Figure 2: An example of an engaging database based on the results yielded by this study; the domain is relatively easily understood and common, and enables the delivery of digital goods; *NT* = 6, *NA* = 27, *NFK* = 6, *COS* = 36, *DRT* = 2, and with a median of 275 rows per table (not visualized here); foreign keys are indicated in blue

- Familiarity
- Connection to practice
- Learning opportunity
- Versatility
- Social and ethical issues
- Practical constraints
- Simplicity

Suggestion 1: generating questions

Table 3: Summary of expert evaluation.

Question	Response	Count	Percentage
1. The exercise description was clear	Yes	273	96.5%
	Partially	10	3.5%
	No	0	0.0%
2. The exercise description matched the selected theme	Yes	272	96.1%
	Partially	7	2.5%
	No	4	1.4%
3. The exercise description matched the selected topic	Yes	270	95.4%
	Partially	9	3.2%
	No	4	1.4%
4. The exercise description matched the selected concept	Yes	248	87.6%
	No	35	12.4%
5. Included concepts that were too advanced	Yes	14	4.9%
	No	269	95.1%
6. The exercise difficulty matched the selected difficulty	Too easy	112	39.6%
	Okay	154	54.4%
	Too difficult	17	6.0%
7. Shallow vs. deep personalization	Deep	75	26.5%
	Unsure	27	9.5%
	Shallow	181	64.0%

Evanfiya Logacheva, Arto Hellas, James Prather, Sami Sarsa, and Juho Leinonen. 2024. Evaluating Contextually Personalized Programming Exercises Created with Generative AI. ICER 2024. https://doi.org/10.1145/3632620.3671103

The prompts themselves are simple, as they should have a positive effect on the time teachers need to generate exams. The first two prompts should generate the basics:

(1) Create a database schema about [domain].
(2) Generate mock data for this database schema.

Then, the schema from question 1 was reused to generate questions:

(3) Create a SQL question testing students' understanding of grouping, using the database schema.
(4) Create a SQL question testing students' understanding of `HAVING`, using the database schema.
(5) Create a SQL question only using filtering and give the answer.
(6) Create a SQL question testing students' understanding of empty-`JOIN`s and give the answer.
(7) Create an interesting SQL question for this database schema and explain chain-of-thought.
(8) Create a hard SQL question for this database schema and explain chain-of-thought.

Willem Aerts, George Fletcher, and Daphne Miedema. 2024. A Feasibility Study on Automated SQL Exercise Generation with ChatGPT-3.5. DataEd 2024. https://doi.org/10.1145/3663649.3664368

Error messages

Table 4. Typical Characteristics of the SQL Error Messages of Eight DBMSs; It is Worth Noting that These are Typical Characteristics Based Only on the 16 Types of Errors Studied

DBMS	Characteristics of SQL error messages
MySQL (with InnoDB)	Sometimes contain error codes at the beginning of the message; both brief and wordy messages; general suggestions to check the manual; line numbers sometimes present; sometimes replicates a part of the query; non-uniform error messages.
Oracle Database	Error codes at the beginning of the message; brief messages; no line numbers; general messages.
PostgreSQL	No error codes; line numbers; specific error position is indicated by a free-standing circumflex; sometimes provides hints; replicates the erroneous line; complete sentences.
SQL Server	Error codes and additional environmental variable information at the beginning of the message; line numbers; a single message may identify multiple errors; replicates the erroneous position.
CockroachDB	No error codes; does not explicitly state that there is an error; both brief and wordy messages; no line numbers; sometimes the specific error position is indicated by a free-standing circumflex; sometimes provides general hints; sometimes replicates the query up to the position of the error.
SingleStore (with InnoDB)	Error codes at the beginning of the message; general suggestions to check the manual; line numbers sometimes present; sometimes replicates a part of the query.
NuoDB	Error codes at the beginning of the message; sometimes replicates parts of the query; sometimes the specific error position is indicated by a free-standing circumflex; sometimes explains what was expected at the erroneous position.
VoltDB	No error codes; no line numbers; replicates the whole query; sometimes provides hints; sometimes explains what was expected at the erroneous position.

Taipalus, T., & Grahn, H. (2023). Framework for SQL error message design: A data-driven approach. ACM Transactions on Software Engineering and Methodology.

Error messages

Table 3. The SQL Error Message Design Framework Consists of Three High-Level Themes Consisting of a Total of Nine Guidelines

Where	**Provide line number**: as accurately as possible, show the user on which line the error is. **Specify the error position**: point to the position of the error on the erroneous line.
What	**Explain what causes the error**: describe what is missing, extraneous, ill-placed, or incorrect. **Explain why the error occurs**: describe what principle is violated. **Place the most important information first**: let the user choose whether to read further.
How	**Provide suggestions on how to fix the error**: use reserved wording, as the intent of the user is unknown. **Provide working examples of similar query concepts**: show how a query concept is used as a part of a query.
Remove unnecessary elements: remove error codes, host names etc., or move them to the end of the message.	
Use plain English: use well-understood terms, or explain complex terms using simple natural language words.	

Table 4. Typical Characteristics of the SQL Error Messages of Eight DBMSs; It is Worth Noting that These are Typical Characteristics Based Only on the 16 Types of Errors Studied

DBMS	Characteristics of SQL error messages
NuoDB	Error codes at the beginning of the message; sometimes replicates parts of the query; sometimes the specific error position is indicated by a free-standing circumflex; sometimes explains what was expected at the erroneous position.
VoltDB	No error codes; no line numbers; replicates the whole query; sometimes provides hints; sometimes explains what was expected at the erroneous position.

Taipalus, T., & Grahn, H. (2023). Framework for SQL error message design: A data-driven approach. ACM Transactions on Software Engineering and Methodology.

Suggestion 2: use the LLM (tutor) to help

Train an agent to be an interactive manual for the Database Management System

Let LLMs generate error message modifications

Let LLMs uncover and share all errors they can detect in one go

Suggestion 3: New assignment types

Prompt challenge: ask students to find natural language questions that the LLM cannot accurately translate to SQL (or some other query language).

Prompt problem: Let students craft natural language texts that lead an LLM to generate the appropriate result.

Assign reflection questions that focus on understanding and metacognition

Explain in Plain English: Give students code and let them write a natural language prompt that results in the same piece of code.

Suggestion 3: New assignment types

Prompt challenge: ask students to find natural language questions that the LLM cannot accurately translate to SQL (or some other query language).

Prompt problem: Let students craft natural language texts that lead an LLM to generate the appropriate result.

Assign reflection questions that focus on understanding and metacognition

Explain in Plain English: Give students code and let them write a natural language prompt that results in the same piece of code.

Debugging exercises: Give students buggy code and let them design test cases to unveil the issues.

Suggestion 4: Amend learning objectives

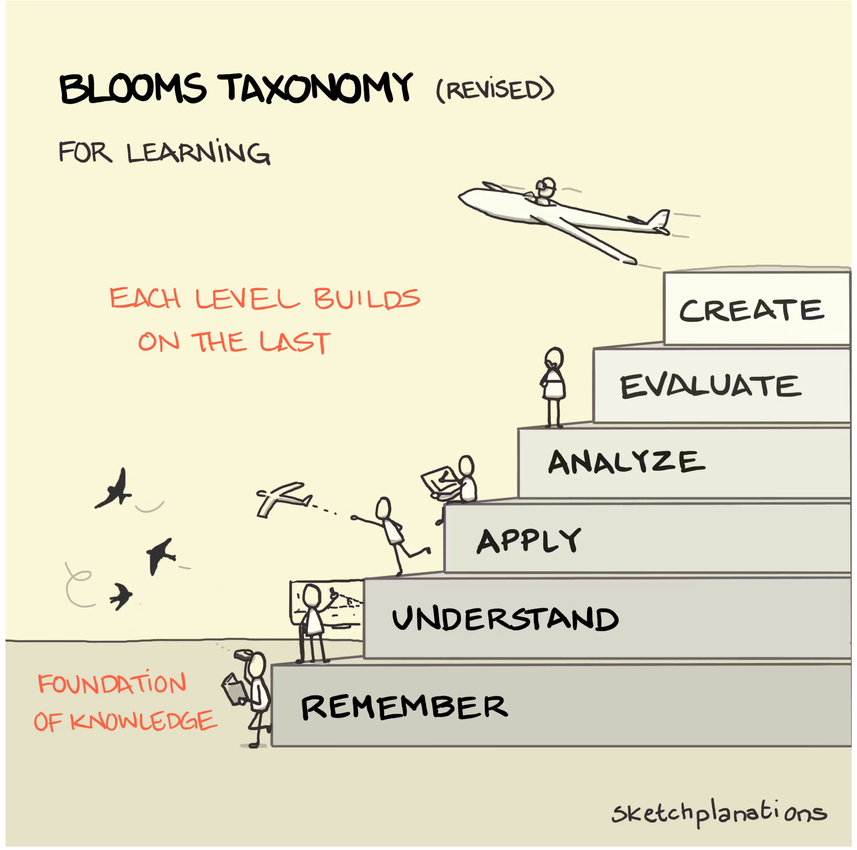

Image: Jono Hey, sketchplanations.com

Suggestion 4: Amend learning objectives

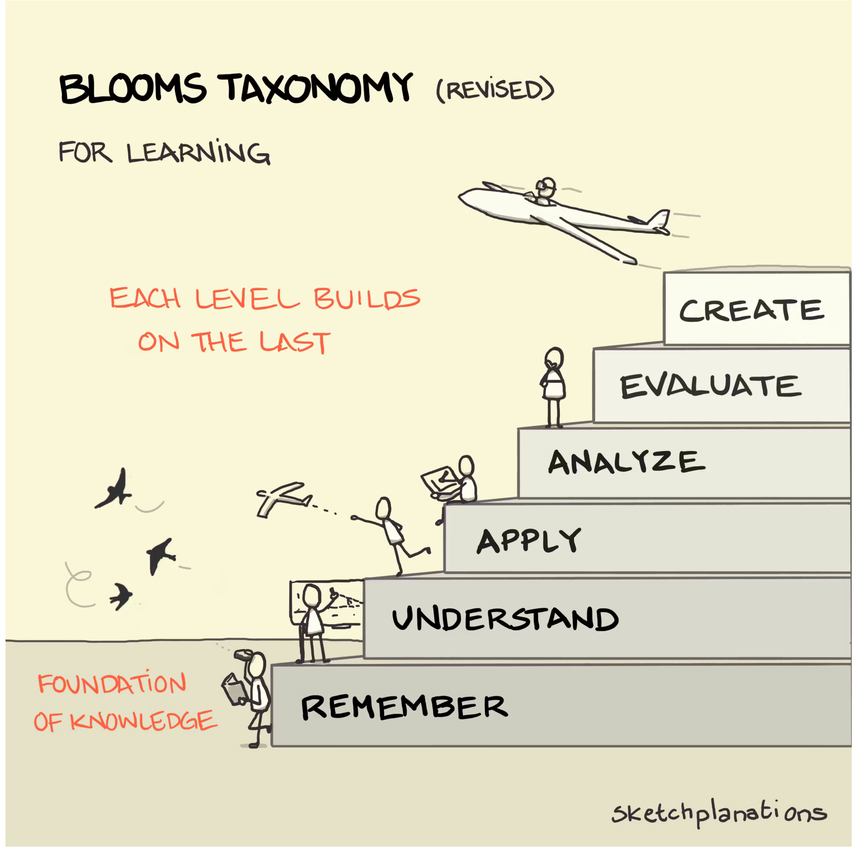

Image: Jono Hey, sketchplanations.com

Dennis J. Bouvier, Bruno Pereira Cipriano, Richard Glassey et al. 2026. The Rest of the Robots: Generative AI in Post-introductory Computing Education. ITiCSE Working Group Reports. https://doi.org/10.1145/3760545.3783970

Suggestions:

- **Critical evaluation and verification:** Detecting, localizing, and correcting GenAI errors; including evaluating and understanding both AI-generated and human-written code.
- **Responsible and effective GenAI use:** Crafting reproducible prompts, documenting provenance, and articulating when and why to trust outputs.
- **Iterative improvement:** Using GenAI for feedback and refinement while maintaining human oversight and standards of evidence - developing enhanced critical thinking and problem-solving capabilities.
- **Design and planning with GenAI:** Leveraging GenAI to expand option sets, surface constraints, and reason about trade-offs.
- **Communication about GenAI-assisted work:** Explaining decisions, limitations, and ethical considerations to technical and non-technical audiences.

Interested in work like this?

Call for participation: DataEd’26 at EDBT/ICDT

Workshop

DataEd: Bridging education practice with education research

Time	Program	Title and presenter(s)
10:30	Coffee	EDBT/ICDT provides coffee in the Auditorium Foyer
10:45	Opening and welcome	Remarks from the DataEd Workshop Chairs.
11:00	Keynote + Q&A	**CSEd keynote:** Title to be announced Prof. Craig Zilles, University of Illinois at Urbana-Champaign
11:45	Paper Session 1 - Data Systems	**Paper 1:** GRANT Privileges, REVOKE Risk: Safe and Scalable Teaching of Database Administration with Isolated Containers Andrzej Wójtowicz (Adam Mickiewicz University in Poznań), Maciej Prill (Adam Mickiewicz University in Poznań)
12:00 PM		**Paper 2:** A Collection of Demonstrations with PostgreSQL for Teaching and Learning Database System Internals Stefan Halfpap(TU Berlin), Daniel Hristov (TU Berlin), Volker Markl (TU Berlin)
12:15		**Paper 3:** SemiStructQuest: Unveiling NoSQL Database Concepts Through Gamified Learning Platform Analyses Nelly Barret (INSA Lyon), Mélina Verger (INSA Lyon)
12:30	Lunch	EDBT/ICDT provides lunch in Tampere Hall (3rd floor)
14:00	Keynote + Q&A	**Data Systems Keynote:** Title to be announced Prof. Azza Abouzied, New York University Abu Dhabi
14:45	Paper Session 2 - Automation	**Paper 4:** AI-Assisted Generation of SQL Comprehension Questions Martin Goodfellow (University of Strathclyde), Alasdair Lambert (University of Strathclyde), Andrew Fagan (University of Strathclyde), Robbie Booth (University of Strathclyde)
15:00		**Paper 5:** A Proposal for Revising SQL Error Taxonomies Based on Automated Detection Davide Ponzini (University of Genoa), Giovanna Guerrini (DIBRIS- University of Genova), Barbara Catania (DIBRIS-University of Genoa)
15:15	Sponsor Talk	Oracle
15:25	Coffee break	EDBT/ICDT provides coffee in the Auditorium Foyer
16:00	Paper session 3 - Modeling	**Paper 6:** Seasoning Data Modeling Education with GARLIC: A Participatory Co-Design Framework Viktoriia Makovska (Ukrainian Catholic University), Ihor Michurin (Kharkiv National University of Radio Electronics), Mariia Tokhtamysh (Kharkiv National University of Radio Electronics), George Fletcher (Eindhoven University of Technology), Julia Stoyanovich (New York University)
16:15		**Paper 7:** Teaching Query-Driven Design in Aggregate-Oriented NoSQL Systems: Resources, Methodology, and Tool Support Barbara Catania (DIBRIS-University of Genoa), Giovanna Guerrini (DIBRIS- University of Genova), Amer Al Khoury (DIBRIS-University of Genova)
16:30	Guided discussion on the future of Data Systems Education	
17:30	Closing	Remarks from the DataEd Workshop Chairs.

DataEd 2026

EDBT/ICDT - Tampere

CONCLUSION

GenAI for data systems education is here

INDE lab
University of Amsterdam
Informatics Institute

Daphne Miedema
daphnemiedema.nl