(Neural ∘ Symbolic ∘ Neural)(Everything)

The Onion Framework for Modern Query Engines

Wenchao Bai

wbai@seu.edu.cn

October 7, 2025

Outline

1 Notes on SIGMOD Panel

2 Mission

3 Neural?

4 Symbolic ∘ Neural

5 Neural ∘ Symbolic ∘ Neural

Background: The Shifted Funding

Eugene Wu et al. [15] reported the shifted funding in US:

- **DB**: $4B NSF expected budget for 2026.
- **AI**: $>$ $100B VC funding in 2024, $>$ $50B in Q1 2025.

Find the New Chalice of Data (1/3): Are We Polishing a Round Ball?

Where Does Academic Database Research Go From Here?

Continuing to improve RDBMS technology is helpful, but not necessarily competitive with industry, and the gains are increasingly marginal – we are **polishing a round ball**.

A promising direction is AI, but while AI is "an application of data," it doesn't seem like AI **needs** us.

— Eugene Wu, ACM SIGMOD Blog[1]

[1] https://wp.sigmod.org/?p=3801

Find the New Chalice of Data (2/3): Directions

Three points along a continuum in decreasing levels of ambition:

1. **Find the north star**: Does this problem matter to the world? And compared to all of industry and academia, is academic database research necessary to solve it?
2. **Constellation of capabilities**: Articulate *Essential Capabilities* that are missing today: without which applications simply do not and cannot otherwise exist.
3. **A sky full of stars**: Evangelize declarative thinking to cultures throughout the world, of which AI is just one culture. We must clarify what data- (or declarative-? scalable-?) thinking means[2].

[2] As emphasized in [15], core database principles is our competitive advantage, such as (1) "independence between physical and logical", (2) "declarativeness", and (3) "automatic scalability".

Find the New Chalice of Data (3/3): What About AI?

Potential Research Direction in AI.

What if we could query anything and everything in the world using LLMs?

- What does query execution and optimization on LLMs look like?
- How can LLMs aid optimizers?
- What if LLMs were access methods?
- ...

But, are we **uniquely positioned** to dominate this problem? (See: Appendix[p. 41])

Positions of DB Researchers

Optimists:

- **Dan Suciu** (UW): "The people who need us, they know where to find us."
- **Joseph M. Hellerstein** (UC Berkeley): "The time spent debating the foolishness could be devoted to far more constructive purposes."[3]

Reformists:

- **Sihem Amer-Yahia** (CNRS): "... to reach out to colleagues in other **disciplines**."
- **Jens Dittrich** (Saarland Univ.): "We should work more on other topics like **usability**, revise interfaces (instead of performance)."

More positions in Appendix[p. 42].

[3] https://jhellerstein.github.io/blog/sigmod-optimism/

Outline

1 Notes on SIGMOD Panel

2 Mission

3 Neural?

4 Symbolic ◦ Neural

5 Neural ◦ Symbolic ◦ Neural

Modern Query Engines

- **Assumption**: Closed-world[4] → Open-world (Generalization capabilities)
- **Capability**: Retrieval → Reasoning (Reasoning and planning capabilities)
- **Data object**: Relational data → Unstructured multi-modal data (Semantic understanding and extraction capabilities)
- **Query language**: SQL → Natural language (Usability and accessibility)

Refer to [4] for more information.

[4] **Closed-world assumption (CWA)**: A statement that is true is also known to be true. Therefore, conversely, what is not currently known to be true, is false.

Outline

1 Notes on SIGMOD Panel

2 Mission

3 Neural?

4 Symbolic ∘ Neural

5 Neural ∘ Symbolic ∘ Neural

Is LLM a Perfect Proxy?

Parameswaran et al. [6] studied prompt engineering from the perspective of declarative crowdsourcing. The core question to answer is:

- We could construct multiple logically equivalent prompts to achieve the same goal, how can we choose / construct the most appropriate one (in terms of accuracy, cost, *etc.*)?

Q1: Varying Prompting Strategies

Task 1: Semantic Ranking.[5]

Given 20 ice-cream flavors, rank them by how "chocolatey" they are.

1. **Baseline**: List all the items in the prompt and ask LLM to rank them directly.
2. **Coarse-grained**: Rate each item and then sort them based on the ratings.
3. **Fine-grained**: Employ $O(N^2)$ pairwise comparisons to rank the items.

Evaluation: Fine-grained approach trades $\sim 100\times$ tokens for 20% improvement in Kendall Tau-β[6] over the baseline.

[5] Other case studies mentioned in this paper [6] are available in Appendix [pp. 45–48].

[6] **Kendall rank correlation coefficient**: a standard metric to compare rankings. See Appendix[pp. 43–44] for more details.

Opportunities for Improvement

- **O1**: How to identify building blocks (*e.g.*, tools, operators, *etc.*) that can rewrite and revise the original prompt?
  - In the previous example, the building blocks contains an enumerator (tool) to generate flavor pairs, a judger (oper ator) that determines the relative ranking of two flavors, and a composer (tool) to sort the flavors based on the pairwise comparisons.
- **O2**: How to optimize the efficiency?
  - The $O(N^2)$ LLM-calls is impractical for large N.
- **O3**: Can we provide theoretical guarantees?
  - The LLM-based approaches cannot guarantee the output quality.

Outline

1 Notes on SIGMOD Panel

2 Mission

3 Neural?

4 Symbolic ◦ Neural

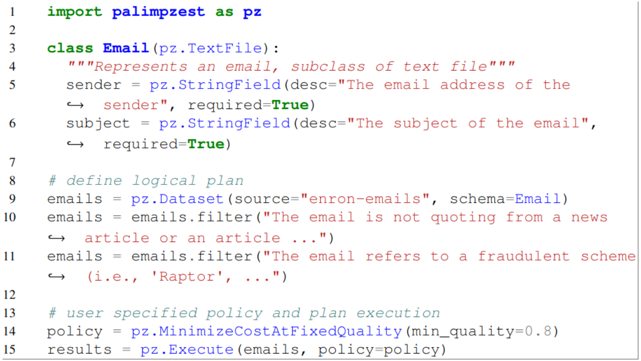
5 Neural ◦ Symbolic ◦ Neural

Symbolic ∘ Neural: Steering LLM-Powered Queries

- **Control the logic flow**: Fine-grained design under declarative frameworks[7].
  - LOTUS [7], Palimpzest [5], UQE [2], ELEET [9]
- **Control the budget**: Reduce LLM calls while maintaining acceptable quality.
  - Model cascade: BARGAIN [16]
  - Approximation: LOTUS [7], UQE [2]
  - Small language models (SLMs): ELEET [9]
  - Cost-based optimization: Palimpzest [5]
- **Control the quality**: Ensure intermediate and final results meet expectations.
  - Human-in-the-loop: ThalamusDB [3]
  - Assertion synthesis: SPADE [8]

[7] **Declarative** means the focus is "what we want" instead of "how to implement". For example, SQL is a typical declarative language whereas C is a typical **imperative** language.

Declarative Frameworks (1/2): Palimpzest

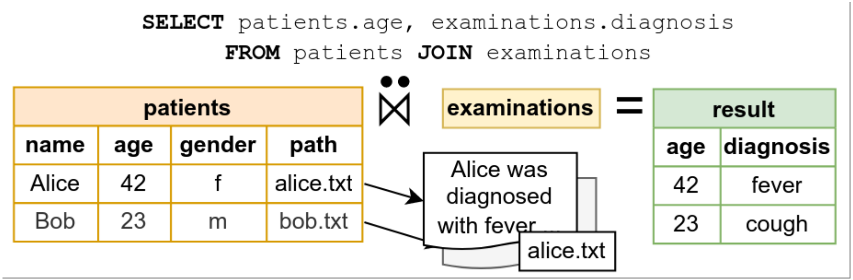


```
1   import palimpzest as pz
2
3   class Email(pz.TextFile):
4     """Represents an email, subclass of text file"""
5     sender = pz.StringField(desc="The email address of the
      ↪  sender", required=True)
6     subject = pz.StringField(desc="The subject of the email",
      ↪  required=True)
7
8   # define logical plan
9   emails = pz.Dataset(source="enron-emails", schema=Email)
10  emails = emails.filter("The email is not quoting from a news
    ↪  article or an article ...")
11  emails = emails.filter("The email refers to a fraudulent scheme
    ↪  (i.e., 'Raptor', ...")
12
13  # user specified policy and plan execution
14  policy = pz.MinimizeCostAtFixedQuality(min_quality=0.8)
15  results = pz.Execute(emails, policy=policy)
```

Figure: Running example of Palimpzest.

1. Define the data schema (line 3-6.)
2. Declare the data source (line 9.)
3. Filter the data using semantic operator “filter” (line 10-11.)
4. Define the execution policy and execute the query (line 14-15.)

Declarative Frameworks (2/2): ELEET



```
SELECT patients.age, examinations.diagnosis
FROM patients JOIN examinations
```

patients			
name	age	gender	path
Alice	42	f	alice.txt
Bob	23	m	bob.txt

result	
age	diagnosis
42	fever
23	cough

Figure: Example of a query that executes a multi-modal join between a patient table and examination reports. ELEET analyzes the texts and extracts values for each queried attribute, such as the diagnosis from each examination report.

Efficiency Optimization (1/5): Model Cascade

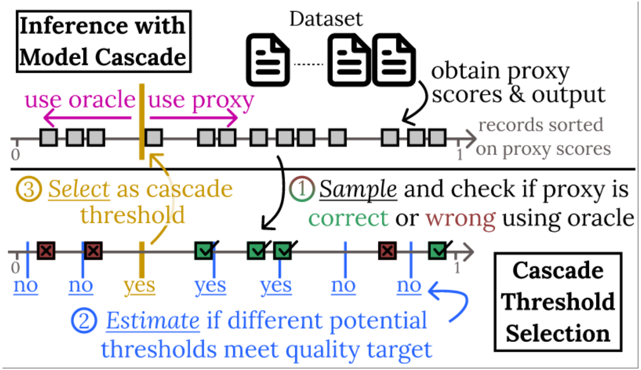

Inference with Model Cascade
Dataset
obtain proxy scores & output
use oracle
use proxy
0
1
records sorted on proxy scores
③ Select as cascade threshold
① Sample and check if proxy is correct or wrong using oracle
0
1
no
no
yes
yes
yes
no
no
② Estimate if different potential thresholds meet quality target
Cascade Threshold Selection

Figure: Overview of model cascade.

- **Intuition/Assumption**: High-confidence output from a small model is likely to be correct.
- BARGAIN [16] provides tight theoretical guarantees through task and data-aware sampling, estimation, and threshold selection.
- **Limitation**: Utility degrades when the proxy model is not well-calibrated [11].

Efficiency Optimization (2/5): Proxy-Based Approximation

- **Intuition**: Use a **fast-but-imperfect** approximate proxy to handle easy cases, reserving the **slow-but-accurate** model only for hard decisions.
- Examples of LOTUS [7]:
  - **Filter**: Use embedding-based classifier or distilled LLMs to filter out obvious matches/mismatches.
  - **Join**: Use embedding-based similarity to filter tuple pairs.

- **Limitations**:
  - Optimization degree is low; cannot optimize at the level of plan structure.
  - Inappropriate adoption of approximation methods results in low accuracy

Efficiency Optimization (3/5): Approximation for Aggregation Queries.

Aggregation Query in UQE.

SELECT COUNT(*) as count
FROM movie_reviews
WHERE "the review is positive";

- **Intuition**: Aggregation queries can be accelerated by reducing the amount of data processed by LLMs.
- UQE [2] adopts **unbiased** stratified sampling for accelerating **aggregation** queries:
  - Embed all rows and cluster them into K groups.
  - Perform stratified sampling within clusters to select a small number of rows.
  - Use weighted averaging of sampled results to unbiasedly estimate aggregation queries

- **Limitation**: This method is not universal, only support aggregation queries.

Efficiency Optimization (4/5): Small Language Models

- **Intuitions**: (1) SLMs are more efficient than LLMs, ensuring efficient online extraction; (2) Information in tables can help locate structured information in text.
- **Limitations**: (1) Cannot support complex semantic analytics; (2) Lack of world knowledge; (3) Impractical assumption: attributes in text are known.

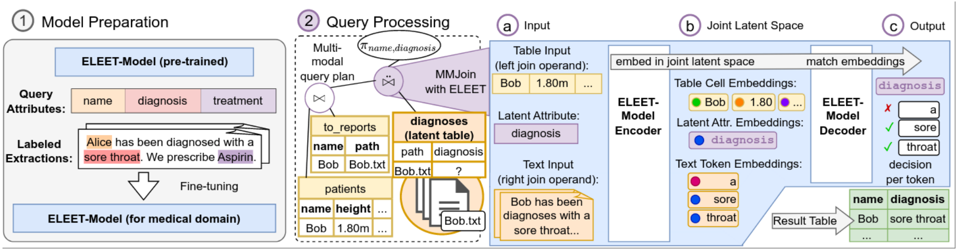


Figure: Overview of ELEET [10].

Efficiency Optimization (5/5): Cost-Based Optimization

- **Intuition/Assumption**: If operators are independent, we can compose operators estimations to estimate plan performance (to avoid exponential searching space.)
- Method of Palimpzest [5]:
  - Executes a set of plans on a small set of **sampled data**.
  - Obtain per-operator estimates (*e.g.*, distribution of runtimes, per-record cost and quality of each operator.)
  - Estimate performance of each plan by composing its per-operator estimates.

- **Limitation**: Estimation by executing over sampled data is time-consuming and inaccurate, which limits optimization effectiveness

Approximate Query Processing by Human-in-the-Loop

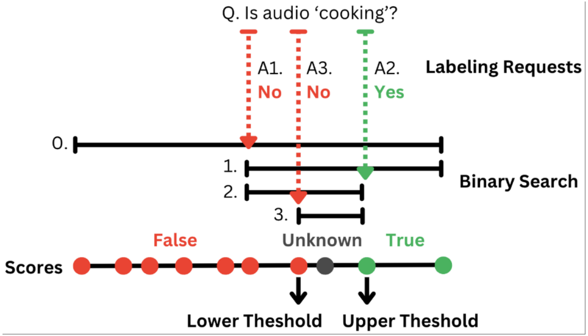


Figure: The labeling process in ThalamusDB [3].

- **Motivation**: Labels are usually unaccessible or expensive to obtain.
- **Intuition**: Users can be seen as "oracles" that provide labels.
- **Objective**: Balance the human cost and the quality of query results.
- **Limitations**: (1) Human effort; (2) Depends on the assumption: the more confident the model is, the more accurate the prediction is.

Assertion Synthesis (1/2): SPADE

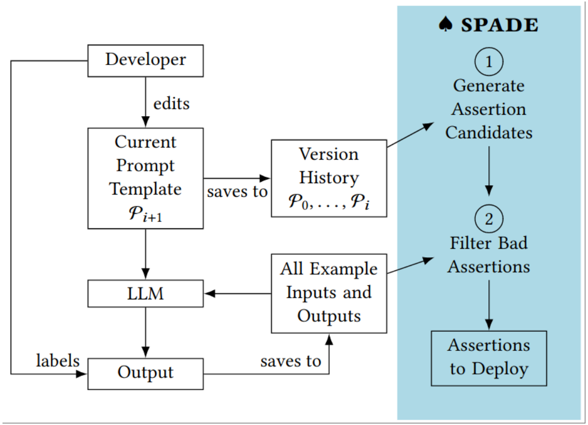
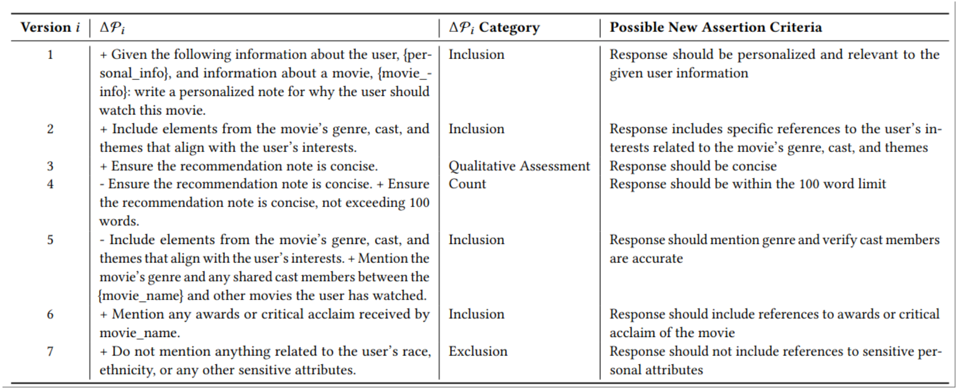

Developer
edits
Current Prompt Template $\mathcal{P}_{i+1}$
saves to
Version History $\mathcal{P}_0, \ldots, \mathcal{P}_i$
LLM
All Example Inputs and Outputs
labels
Output
saves to
♠ SPADE
1
Generate Assertion Candidates
2
Filter Bad Assertions
Assertions to Deploy

Figure: Workflow of SPADE.

- **Motivation**: Monitor the data quality through LLM pipelins.
- **Intuition**: We can mine prompt version histories to identify assertion criteria for LLM pipelines [8].
- **Objective**: Minimal set of assertions with qualified coverage and False Failure Rate.

- **Limitations**: (1) Only focus on single-prompt pipelines; (2) lack of labeled data; (3) LLM dependencies.

Assertion Synthesis (2/2): Delta-Driven Assertion Synthesis



Version i	$\Delta\mathcal{P}_i$	$\Delta\mathcal{P}_i$ Category	Possible New Assertion Criteria
1	+ Given the following information about the user, {personal_info}, and information about a movie, {movie_info}: write a personalized note for why the user should watch this movie.	Inclusion	Response should be personalized and relevant to the given user information
2	+ Include elements from the movie's genre, cast, and themes that align with the user's interests.	Inclusion	Response includes specific references to the user's interests related to the movie's genre, cast, and themes
3	+ Ensure the recommendation note is concise.	Qualitative Assessment	Response should be concise
4	- Ensure the recommendation note is concise. + Ensure the recommendation note is concise, not exceeding 100 words.	Count	Response should be within the 100 word limit
5	- Include elements from the movie's genre, cast, and themes that align with the user's interests. + Mention the movie's genre and any shared cast members between the {movie_name} and other movies the user has watched.	Inclusion	Response should mention genre and verify cast members are accurate
6	+ Mention any awards or critical acclaim received by movie_name.	Inclusion	Response should include references to awards or critical acclaim of the movie
7	+ Do not mention anything related to the user's race, ethnicity, or any other sensitive attributes.	Exclusion	Response should not include references to sensitive personal attributes

Figure: Comparison of 7 prompt versions for an LLM pipeline to write personalized movie recommendations.

Takeaways

- The declarative framework makes it possible to control the logic flow and optimize both efficiency and effectiveness.
- Operator optimization (model cascading, approximation, SLM, cost-based optimization) can effectively improve query efficiency.
- Human-in-the-loop method and assertion synthesis can effectively control the quality of query results.

Caveat: These approaches focus on operators instead of the **workflow** and require user-specified logic flow which introduce additional **human cost**.

Outline

1 Notes on SIGMOD Panel

2 Mission

3 Neural?

4 Symbolic ∘ Neural

5 Neural ∘ Symbolic ∘ Neural

Neural ◦ Symbolic ◦ Neural: Manipulate Symbols Automatically

To improve the accessibility, several challenges should be addressed:

- **How to align the natural language with the query language?**
  - NL2SQL: TAG [1] (Tabular data), CAESURA [10] (Multi-modal data)
  - NL to self-defined operators: Unify [12], iDataLake [14], AOP [13]
  - Planning & tool usage: CAESURA [10], AOP [13]
- **How to optimize the efficiency of the generated query plan?**
  - Independent parallelism: AOP [13]
  - Cost-based optimization: Unify [12]
  - Fault tolerance: iDataLake [14]

Convert Natural Language to Query Languages

- **Objective**: Find a convertion function f: (NL Query, Operators, Tools, *etc.*) → Query Language, that is logically correct.
- **Solution 1**: Instruct LLMs to generate the query plan.
  - Straightforward yet could be inaccurate (depends on LLM capabilities).
- **Solution 2**: Progressive matching [12].
  - Pros: More robustness and exaplainable.
  - Cons: Inflexibility of operaotrs.

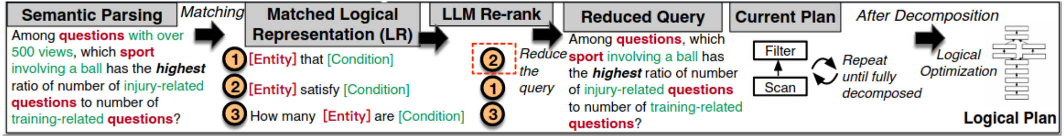

Semantic Parsing
Among questions with over 500 views, which sport involving a ball has the highest ratio of number of injury-related questions to number of training-related questions?
Matching
Matched Logical Representation (LR)
1 [Entity] that [Condition]
2 [Entity] satisfy [Condition]
3 How many [Entity] are [Condition]
LLM Re-rank
2
1
3
Reduce the query
Reduced Query
Among questions, which sport involving a ball has the highest ratio of number of injury-related questions to number of training-related questions?
Current Plan
Filter
Scan
Repeat until fully decomposed
After Decomposition
Logical Optimization
Logical Plan

Figure: Logical plan generation in Unify via progressive matching.

Optimize the Execution Plan (1/3): Independent Parallelism

- **Intuition**: (1) Identifying and parallelizing independent operations can significantly reduce execution time.
- **Method of AOP** [13]: (1) Instruct LLMs to generate pipelines; (2) Optimize pipelines into DAG; (3) Combine different pipelines; (4) Layer-wise execution.

- **Limitation**: The quality of the generated plans rely on LLM capabilities.

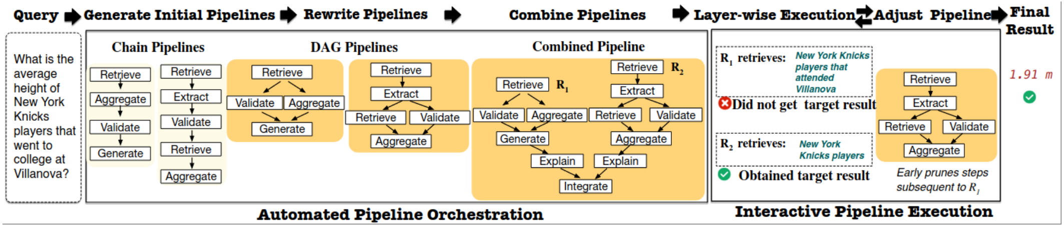

Query
Generate Initial Pipelines
Rewrite Pipelines
Combine Pipelines
Layer-wise Execution
Adjust Pipeline
Final Result
What is the average height of New York Knicks players that went to college at Villanova?
Chain Pipelines
DAG Pipelines
Combined Pipeline
Retrieve
Aggregate
Validate
Generate
Extract
Explain
Integrate
R_1
R_2
R_1 retrieves: New York Knicks players that attended Villanova
Did not get target result
R_2 retrieves: New York Knicks players
Obtained target result
Early prunes steps subsequent to R_1
1.91 m
Automated Pipeline Orchestration
Interactive Pipeline Execution

Figure: Framework of AOP.

Optimize the Execution Plan (2/3): Cost-Based Optimization

- **Observation**: Data points satisfying the query often have high semantic relevance with the query.
- **Mathod of Unify** [12]: (1) Embed records; (2) Retrieve query-related record samples via importance sampling; (3) Estimate the cardinality of query results based on the samples; (4) Optimize the execution plan based on the estimates.

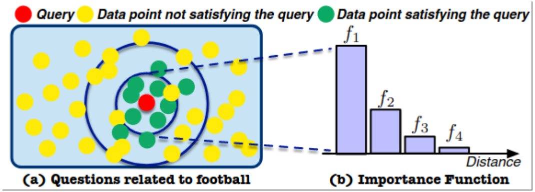

● Query ● Data point not satisfying the query ● Data point satisfying the query
f_1
f_2
f_3
f_4
Distance
(a) Questions related to football
(b) Importance Function

Figure: Importance sampling in Unify.

Optimize the Execution Plan (3/3): Online Plan Adjustment

- **Intuition**: When the execution fails, it's usually beneficial to restore from the checkpoint and switch to other other low-cost plans rather than starting over.

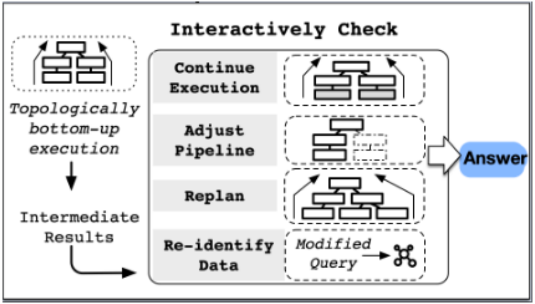


Figure: Pipeline execution in iDataLake [14].

Case Study: Spec-Driven Development with Claude Code

- **Declarativeness**: The development is steered by the qualified task list[8]. Tools and MCP servers can be flexibly plugged in.
- **Efficiency**: Claude Code supports parallel sessions [9].
- **Human effort**: Claude Code supports YOLO (You-Only-Live-Once) mode[10].
- **Quality Control**: (1) Test-driven development; (2) Human-in-the-loop checking.

[8] The spec and the task list can be generated by github/spec-kit.

[9] See: Run Parallel Claude Code Sessions with Git Worktrees.

[10] See: Claude Code best practices.

Takeaways

- Directly instructing LLMs to generate pipeline achieves **limited accuracy**.
- Progressively matching appropriate operators is limited by **inflexibility** of operaotrs, strict requirement of intput/output relationship of operators.
- Optimization strategies: (1) Parallelizing independent operations; (2) Cost-based optimizations; and (3) Online plan adjustment.

Discussion: Where can we go from here?

References I

- Asim Biswal, Liana Patel, Siddarth Jha, Amog Kamsetty, Shu Liu, Joseph E Gonzalez, Carlos Guestrin, and Matei Zaharia.
  Text2sql is not enough: Unifying ai and databases with tag.
  In CIDR, 2025.

- Hanjun Dai, Bethany Wang, Xingchen Wan, Bo Dai, Sherry Yang, Azade Nova, Pengcheng Yin, Mangpo Phothilimthana, Charles Sutton, and Dale Schuurmans.
  Uqe: A query engine for unstructured databases.
  NeurIPS, 37:29807–29838, 2024.

- Saehan Jo and Immanuel Trummer.
  Thalamusdb: Approximate query processing on multi-modal data.
  PACMMOD, 2(3), May 2024.

References II

Guoliang Li, Jiayi Wang, Chenyang Zhang, and Jiannan Wang.
Data+ai: Llm4data and data4llm.
In SIGMOD Tutorials, page 837–843, 2025.

Chunwei Liu, Matthew Russo, Michael Cafarella, Lei Cao, Peter Baile Chen, Zui Chen, Michael Franklin, Tim Kraska, Samuel Madden, Rana Shahout, et al.
Palimpzest: Optimizing ai-powered analytics with declarative query processing.
In CIDR, 2025.

Aditya Parameswaran, Shreya Shankar, Parth Asawa, Naman Jain, and Yujie Wang.
Revisiting prompt engineering via declarative crowdsourcing.
In CIDR, 2024.

References III

Liana Patel, Siddharth Jha, Melissa Pan, Harshit Gupta, Parth Asawa, Carlos Guestrin, and Matei Zaharia.
Semantic operators: A declarative model for rich, ai-based data processing.
PVLDB, page 4171–4184, 2025.

Shreya Shankar, Haotian Li, Parth Asawa, Madelon Hulsebos, Yiming Lin, JD Zamfirescu-Pereira, Harrison Chase, Will Fu-Hinthorn, Aditya G Parameswaran, and Eugene Wu.
Spade: Synthesizing data quality assertions for large language model pipelines.
PVLDB, 17(12):4173–4186, 2024.

References IV

Matthias Urban and Carsten Binnig.
Eleet: Efficient learned query execution over text and tables.
PVLDB, 17(13):4867–4880, 2024.

Matthias Urban and Carsten Binnig.
Caesura: language models as multi-modal query planners.
In CIDR, 2025.

Cheng Wang.
Calibration in deep learning: A survey of the state-of-the-art, 2025.

Jiayi Wang and Jianhua Feng.
Unify: An unstructured data analytics system.
In ICDE, pages 4662–4674, 2025.

References V

Jiayi Wang and Guoliang Li.
Aop: Automated and interactive llm pipeline orchestration for answering complex queries.
In CIDR, 2025.

Jiayi Wang, Guoliang Li, and Jianhua Feng.
idatalake: An llm-powered analytics system on data lakes.
Data Engineering, page 57, 2025.

Eugene Wu and Raul Castro Fernandez.
Where does academic database research go from here?, 2025.

References VI

Sepanta Zeighami, Shreya Shankar, and Aditya Parameswaran.
Cut costs, not accuracy: Llm-powered data processing with guarantees.
PACMMOD, 2026.

A. Does DB Research Dominate LLM-Querying?

TL;DR: Yes and No.

- **Our advantages**. Declarativeness, cost-based optimization, approximation guarantees, and performance optimization. They can do RAG but we can do it better.
- **Poor predicatability**. LLMs as a compute substrate evolve weekly, meaning that any optimization rules based on yesterday's trade-offs are immediately slower, more expensive, or lower quality. (We can only follow LLM companies.)
- **Impossible to reason about**. The responses change for no reason, their tunable parameter is "any text." (Common task with the AI community.)
- **Undefined correctness**. There are no semantics to create correctness benchmarks, only vibes. (Common task with the AI community.)

Similar reasoning should be applied to data preprocessing for AI, video querying, provenance for AI, NL2SQL, ML training and serving, prompt engineering, etc.

B. More Positions for the Panel

- **Pınar Tözün** (IT Univ. of Copenhagen): "I think there are 'Systems for ML' aspects that our community is well-positioned to take on."
- **Leilani Battle** (UW): "(Consider) Not just 'what can we do?' but also 'what should we do?'(*e.g.*, ethical issues, software abusiveness, etc.)"
- **Aditya Parameswaran** (UC Berkeley): "(We need to) fully embrace LLMs as a means to solve the AI-complete problems, *e.g.*, data integration, data cleaning."
- **Sudeepa Roy** (Duke Univ.): "(We should) be open to embrace new ideas and take the opportunity to start new collaborations in the fast evolving landscape of AI."
- **Samuel Madden** (MIT): "Our community still has a role to play in the new AI era."
- **Felix Naumann** (HPI): "The traditional problem space of all things data management is alive and kicking."
- **Paolo Papotti** (EURECOM): "One risk is the limited openness to new topics."
- **James Cowling** (Convex): "The huge gift academia has is the ability to investigate impractical ideas."

C-1. Kendall Tau-β: Intuition

- The Kendall Tau-β coefficient measures the **rank correlation** between two variables X and Y.
- It quantifies how consistently two orderings (e.g., human vs. model) **agree on pairwise comparisons**.
- For any two items (x_i, y_i) and (x_j, y_j):
  - **Concordant**: $(x_i - x_j)(y_i - y_j) > 0$
  - **Discordant**: $(x_i - x_j)(y_i - y_j) < 0$
  - **Tie**: $x_i = x_j$ or $y_i = y_j$
- τ_β represents the **probability difference** between concordant and discordant pairs:

$$\tau_\beta = P(\text{concordant}) - P(\text{discordant})$$

C-2. Kendall Tau-β: Notation and Formula

- Let:

$$\begin{aligned} n_c &: \text{number of concordant pairs} \\ n_d &: \text{number of discordant pairs} \\ n_1 &: \text{pairs tied only on } X \\ n_2 &: \text{pairs tied only on } Y \end{aligned}$$

- The Kendall Tau-β coefficient is defined as:

$$\tau_\beta = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_1)(n_c + n_d + n_2)}}$$

- **Range:** $-1 \leq \tau_\beta \leq 1$
  - $\tau_\beta = 1$: perfect agreement (identical ranking)
  - $\tau_\beta = -1$: perfect disagreement (reverse ranking)
  - $\tau_\beta = 0$: no correlation

D-1. Q2: Hybrid Coarse → Fine-Grained Prompting

Task 2: Alphabetical Sorting.[11]

Given a list of 100 random English words, sort them in alphabetical order.

1. **Baseline**: List all the items in the prompt and ask LLM to sort them directly.
2. **Hybrid strategy**: (a) Ask the LLM to sort the entire list; (b) Drop all hallucinated words; (c) Reinsert the missing words.

Evaluation: The hybrid strategy trades $O(kN)$ additional LLM calls (for k missing words and N partially sorted words) to eliminate hallucinated and missing words.

[11] This example illustrates error mitigation in fine-grained approaches, though alphabetical sorting represents a deterministic task where conventional algorithms would be more appropriate than LLMs.

D-2. Q3: Ensuring Internal Consistency

Task 3: Entity Resolution.

Identify all same citations on the DBLP-Google Scholar dataset.

1. **Baseline**: Employ $O(N^2)$ pairwise LLM comparisons for entity resolution.
2. k-**NN**: (a) Retrieve k nearest neighbors per citation; (b) Execute batch identification within neighborhood clusters for each citation pair ($2k + 2$ citations per LLM-call); (c) Compute transitive closure[12] over identified equivalence pairs.

Evaluation: The baseline exhibits high precision (95.2%) but limited recall (50.3%). Applying transitive closure with 2-NN improves recall to 59.3% while maintaining acceptable precision (92.3%).

[12] *e.g.*, the transitive closure of $\{(1,2),(2,3)\}$ is $\{(1,2),(2,3),(1,3)\}$

D-3. Q4: Leveraging LLM and non-LLM Approaches

Missing Value Imputation.

Given an entity with j attributes $A = \{a_1, \ldots, a_j\}$ and values $E = \{e_1, \ldots, e_j\}$ where e_j is missing, predict e_j based on the known values.

1. **Baseline**: Employ LLM to predict e_j based on A and E directly.
2. k-**NN**[13]: (a) Retrieve k nearest neighbors of the entity; (b) Impute by k-NN if all neighbors have the same value for a_j; (c) Otherwise, impute by the LLM.

Evaluation: (a) The baseline suffers from the misalignment with ground truth (*e.g.*, "Elgato Systems" instead of "Elgato";) (b) The hybrid approach (LLM+k-NN) reduces token consumption by 50% with acceptable accuracy degradation (92.31%→87.69%.)

[13] Besides non-LLM methods, model cascade [16] is an alternative yet more flexible solution.

D-4. Takeaways

Q1: Rather than trying to accomplish the entire objective via a single task, it is beneficial to explore other task types, especially to maximize accuracy.

Q2: Employing hybrid strategies, with coarse-grained tasks first, followed by fine-grained ones, can lead to low cost and high accuracy overall.

Q3: Fixing erroneous LLM responses based on evidence from other responses can be an effective way to improve accuracy.

Q4: Leveraging a non-LLM proxy can help substantially reduce costs while keeping accuracy similar.

Insight: Identifying and reconstructing building blocks of LLM-powered query tasks yields benefits in both effectiveness and efficiency.