EcoLLM: Energy-Aware Benchmarking of LLMs for Data Processing Workloads

aiDM@SIGMOD 2026
Bengaluru



Pratyush Agnihotr Manisha Luthra Carsten Binnig

dfki ai
TECHNISCHE UNIVERSITÄT DARMSTADT
RUHR UNIVERSITÄT BOCHUM
RUB

LLMs are becoming part of data-system execution.

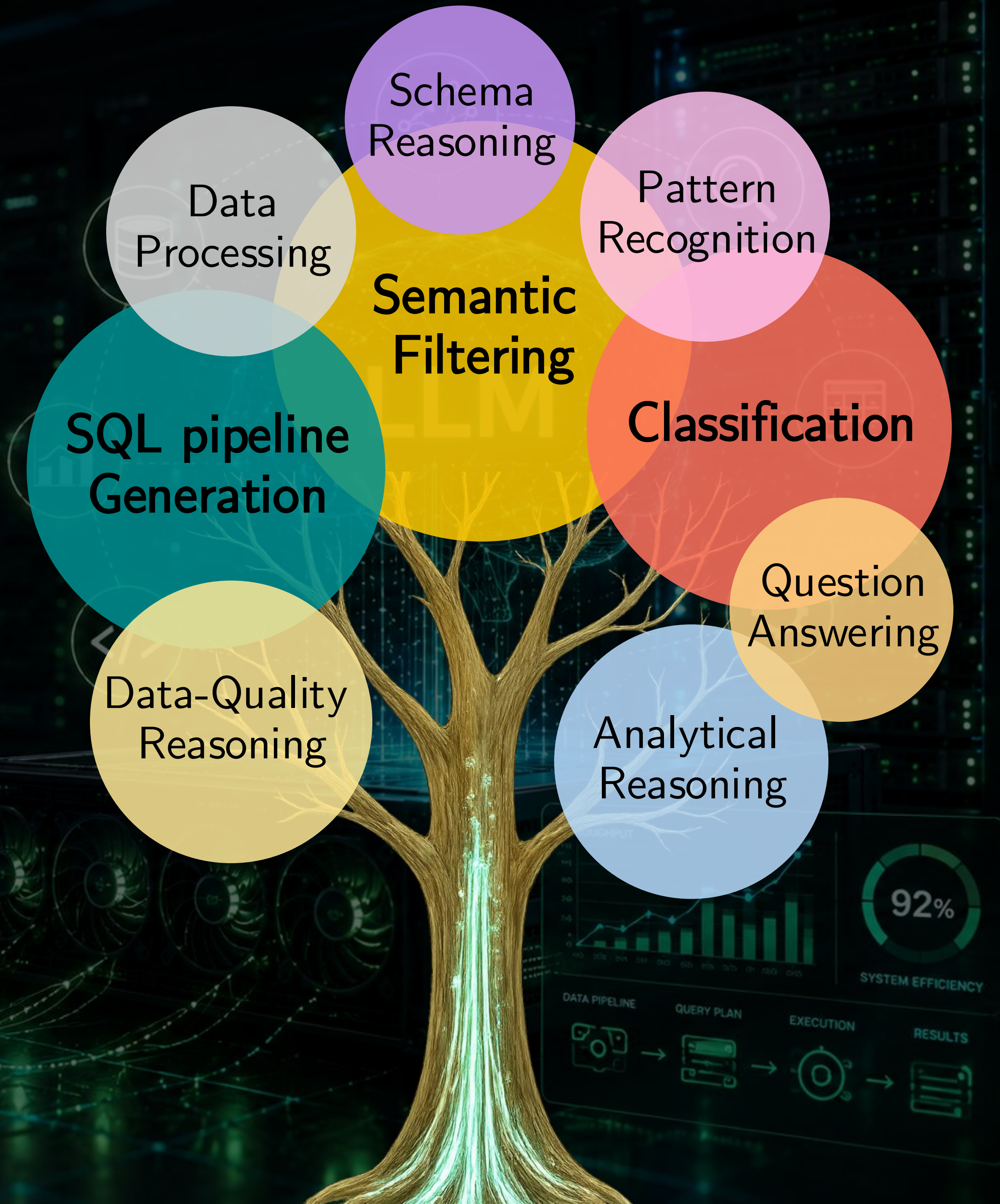
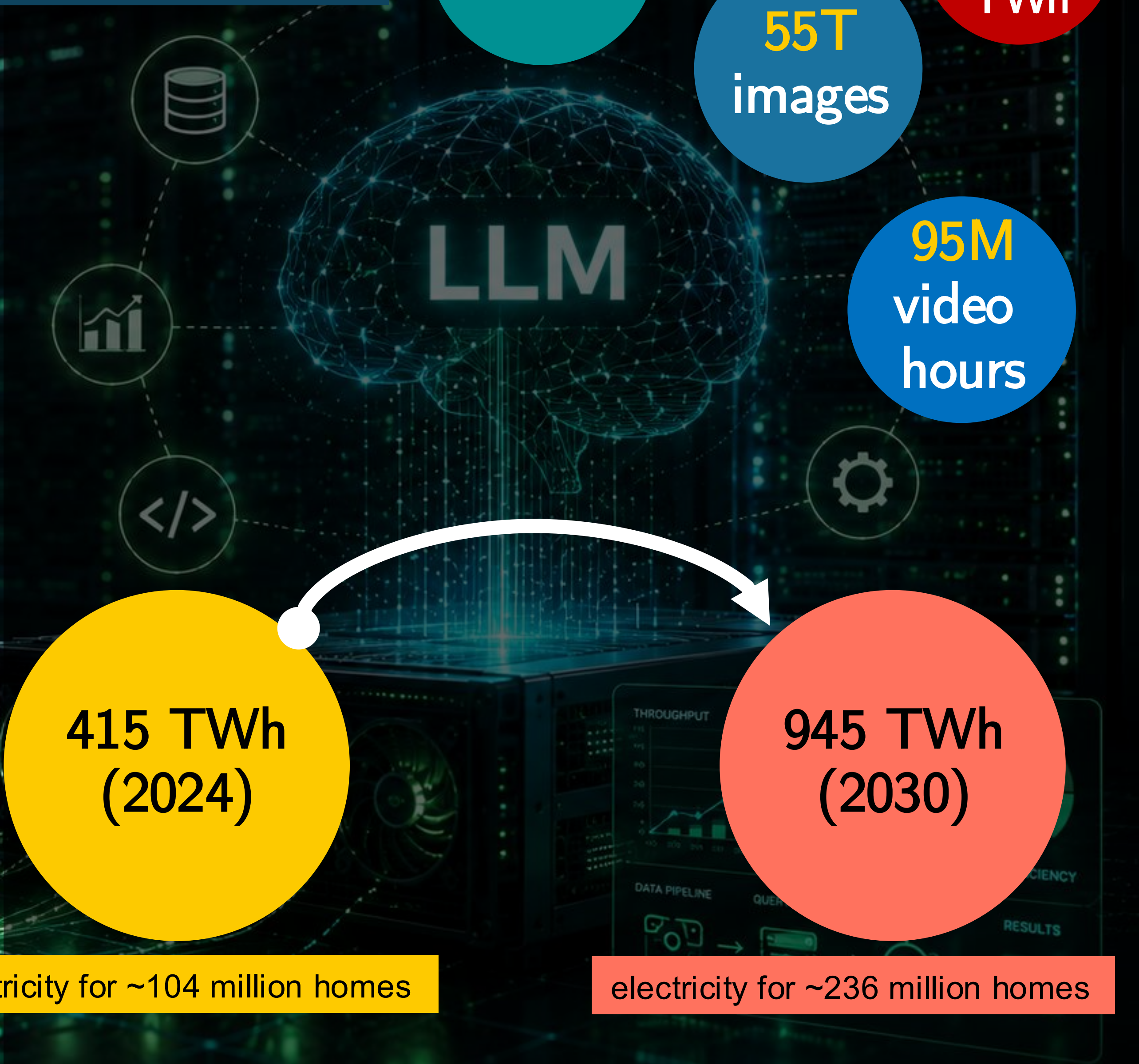


Image source: Google and Dall.e AI generated

The Hidden Energy Cost of Generative AI

LLMs are becoming part of data-system execution.

IEA projects very high global data-centre electricity consumption!



4,250T words

55T images

95M video hours

100 TWh

415 TWh (2024)

electricity for ~104 million homes

945 TWh (2030)

electricity for ~236 million homes

Image source: Google and Dall.e AI generated

The Hidden Energy Cost of Generative AI

4,250T words

55T images

95M video

100 TWh

LLM

LLMs are becoming part of

AI inference is **no longer a ML problem**,
it is becoming a
data systems and infrastructure problem.

IEA reports global data-centre electricity consumption

415 TWh (2024)

945 TWh (2030)

electricity for ~104 million homes

electricity for ~236 million homes

Image source: Google and Dall.e AI generated

Data Systems Optimize Everything Except Energy

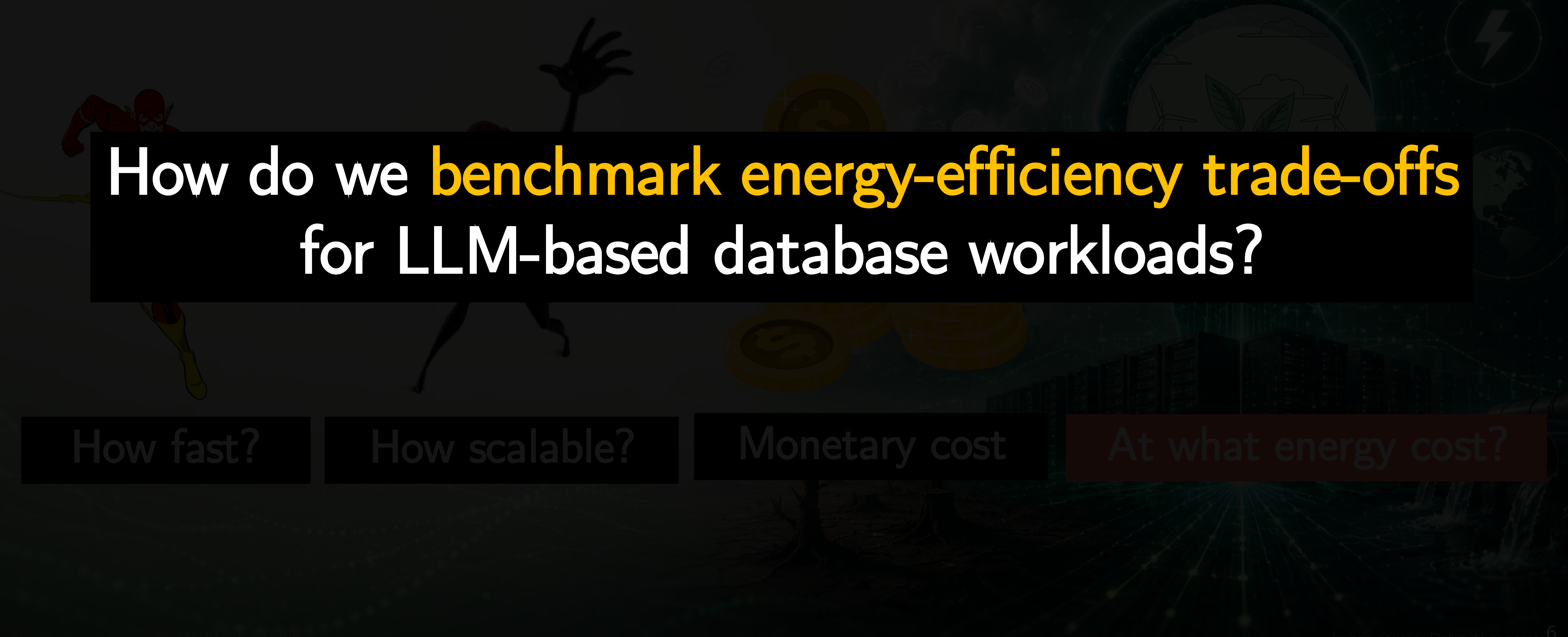
Yet data systems research still asks:

How fast?	How scalable?	Monetary cost	At what energy cost?

Image source: Google and Dall.e AI generated

Data systems optimize everything except energy

Yet data systems research still asks:



How do we benchmark energy-efficiency trade-offs for LLM-based database workloads?

How fast? | How scalable? | Monetary cost | At what energy cost?

Image source: Dall.e AI generated

Introducing EcoLLM

- an energy- and workload-centric benchmark – **Energy is a first-class metric**
- Key insight: **how LLMs based database workloads behave for different workload level parameters and what is their energy-efficiency trade-offs.**

Existing work

NLP Query: What was the average energy consumption in Texas?

Filter → Join → Aggr.

Traditional Query Pipeline

EcoLLM

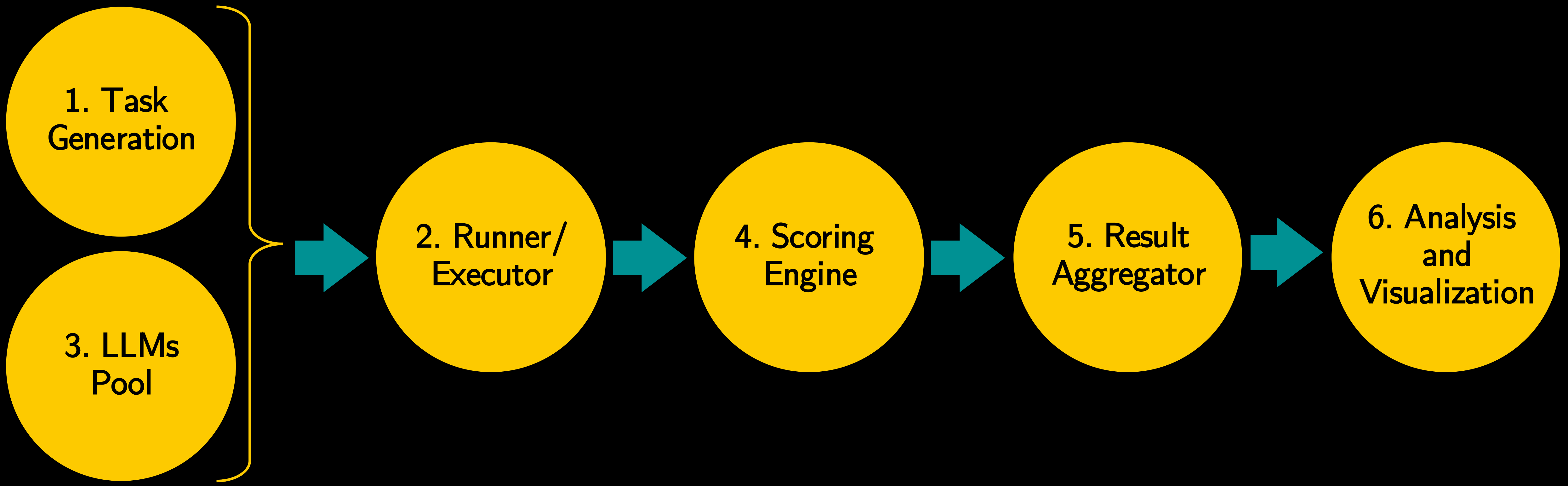
Filter → LLM (Semantic operator) → Aggr.

LLM-Augmented Pipeline

Pipeline construction:
Generate a SQL query to compute the top-2 states by average consumption.

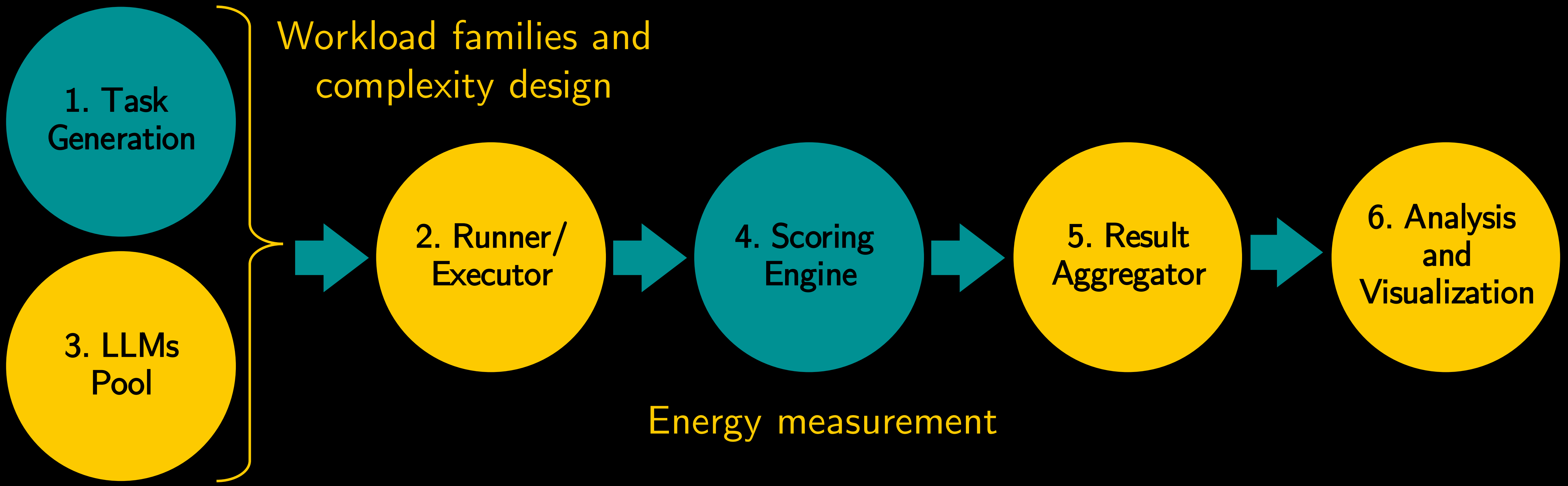
Introducing EcoLLM

- an energy- and workload-centric benchmark – **Energy is a first-class metric**
- **Key insight:** how LLMs based database workloads behave for different workload level parameters and what is their energy-efficiency trade-offs.



Introducing EcoLLM

- an energy- and workload-centric benchmark – **Energy is a first-class metric**
- **Key insight: how LLMs based database workloads behave for different workload level parameters and what is their energy-efficiency trade-offs.**



EcoLLM: Energy Measurement

Not a universal energy meter.

- Energy measurement depends on deployment mode.
- EcoLLM uses a **hybrid strategy**. Different deployment need different measurement methodology.

CODE CARBON

Local models
↓
Resource monitoring
↓
CPU/GPU power, utilization, memory usage, runtime
↓
Energy estimation ≈ **Power × execution time (kWh)**

Hybrid approach

Remote API models
↓
No hardware visibility
↓
Provider-side attribution
↓
Tokens, model type, provider efficiency profile
↓
Energy estimation ≈ tokens × model-specific energy-per-token **(kWh)**

EcoLogits

EcoLLM: Workload Families

- **Family 1: Tabular reasoning**

  Task: Find states where reports indicate increased demand..

- **Family 4: Pipeline code generation**

  Task: Generate a SQL query to compute the top-2 states by average consumption.

LLM executes the operation

LLM generates the program

State	Year	Report Text	Consumption
TX	2020	Demand surged	1200
CA	2020	Industrial growth	950
..	..	..	..

Output = final answer

Output = executable SQL code

LLM Output

```
{
"state": ["TX"]
}
```

```
{"query": "SELECT state, AVG(consumption)
           FROM energy
           GROUP BY state
           ORDER BY AVG(consumption) DESC
           LIMIT 2;"
}
```

EcoLLM: Workload Families

- **Family 1: Tabular reasoning**

  Task: Find states where reports indicate increased demand..

- **Family 4: Pipeline code generation**

  Task: Generate a SQL query to compute the top-2 states by average consumption.

LLM executes the operation

LLM generates the program

Workload families defines type of computation, but do not capture execution difficulty

Output = final answer

Output = executable SQL code

LLM Output

```
{
"state": ["TX"]
}
```

```
{"query": "SELECT state, AVG(consumption)
           FROM energy
           GROUP BY state
           ORDER BY AVG(consumption) DESC
           LIMIT 2;"
}
```

EcoLLM: Workload Families and Complexity Design

- **Family 4: Pipeline code generation**

Task: Generate a SQL query to compute the top-2 states by average consumption.

↓

LLM Output executable SQL code: **Complexity design**

C1: Single-step operation

Table
↓
AVG()

C2:Multi-operator query

Table
↓
Filter
↓
AVG()

C3: Conditional reasoning

Table
↓
GROUP BY
↓
Aggregation
↓
Sorting

C4: Composed analytical pipeline

Energy Table 1+ Table 2
↓
JOIN
↓
FILTER
↓
Aggregation
↓
Ranking
↓
TOP-K

EcoLLM: Workload Families and Complexity Design

- **Family 4: Pipeline code generation**

Task: Generate a SQL query to compute the top-2 states by average consumption.

↓

LLM Output executable SQL code: **Complexity design**

C1: Single-step operation

Table
↓
AVG()

C2: Multi-operator query

Table
↓
Filter
↓
AVG()

C3: Conditional reasoning

Table
↓
GROUP BY
↓
Aggregation
↓
Sorting

C4: Composed analytical pipeline

Energy Table 1+ Table 2
↓
JOIN
↓
FILTER
↓
Aggregation
↓
Ranking
↓
TOP-K

Expected Energy Consumption

Low → C1 → C2 → C3 → C4 High

EcoLLM: Workload Families and Complexity Design

Workload families: What the LLM does.

Complexity: How much reasoning and execution effort is required.

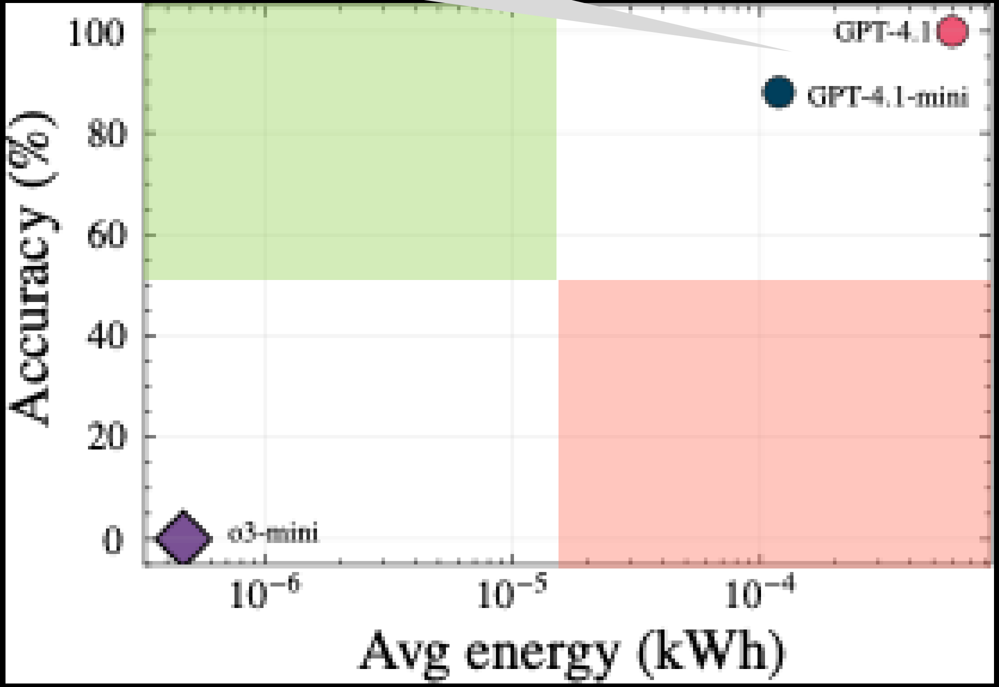
More computation: consumes more energy.

Expected behaviour

EcoLLM Benchmarking Results

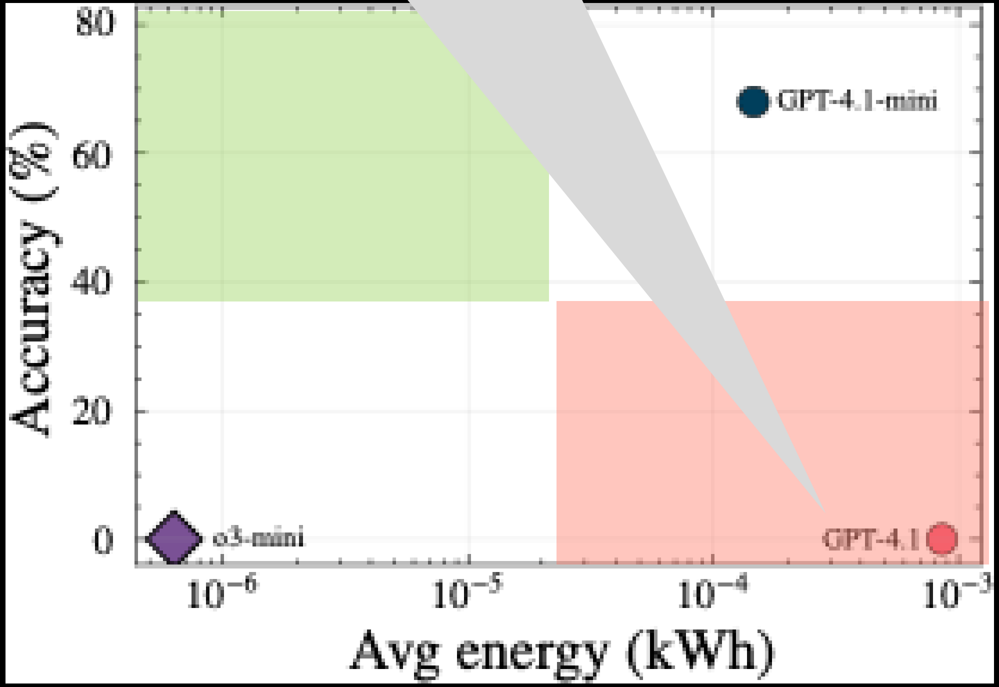
EQ1: Energy vs Performance - Does models trade accuracy for energy efficiency for different workload family?

Higher energy generally improves semantic reasoning validity



Family 1: Tabular reasoning

Higher energy **does not guarantee correctness**



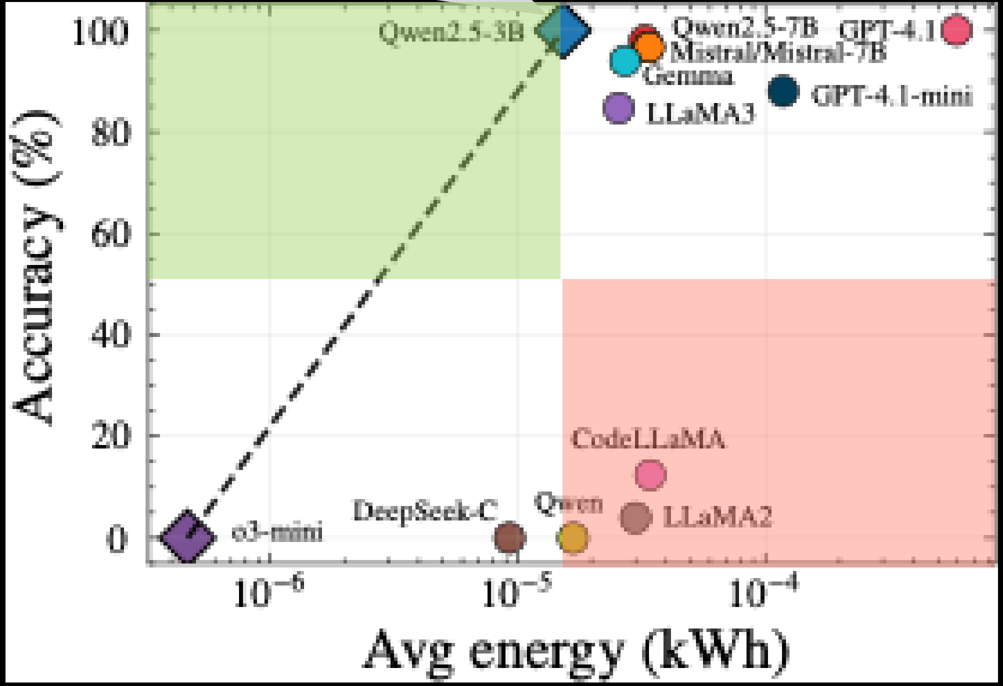
Family 4: Pipeline code generation

EcoLLM Benchmarking Results

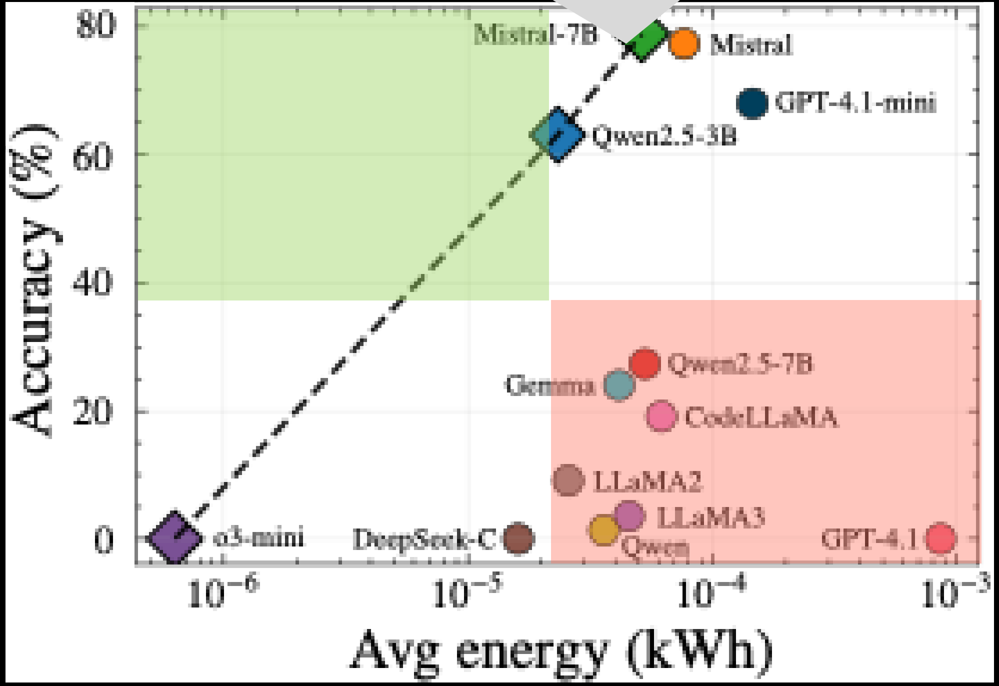
EQ1: Energy vs Performance - Does models trade accuracy for energy efficiency for different workload family?

Smaller models offer better performance with low energy

Specialized models dominate pipeline generation



Family 1: Tabular reasoning



Family 4: Pipeline code generation

EcoLLM Benchmarking Results

EQ1: Energy vs Performance - How do models trade accuracy for energy efficiency for different workload family?

Smaller models offer

b

Specialized models dominate

Larger and energy-intensive models are not always the best choice

For data processing workload, smaller or specialized models can deliver better effectiveness at lower energy cost

Family 1: Tabular reasoning

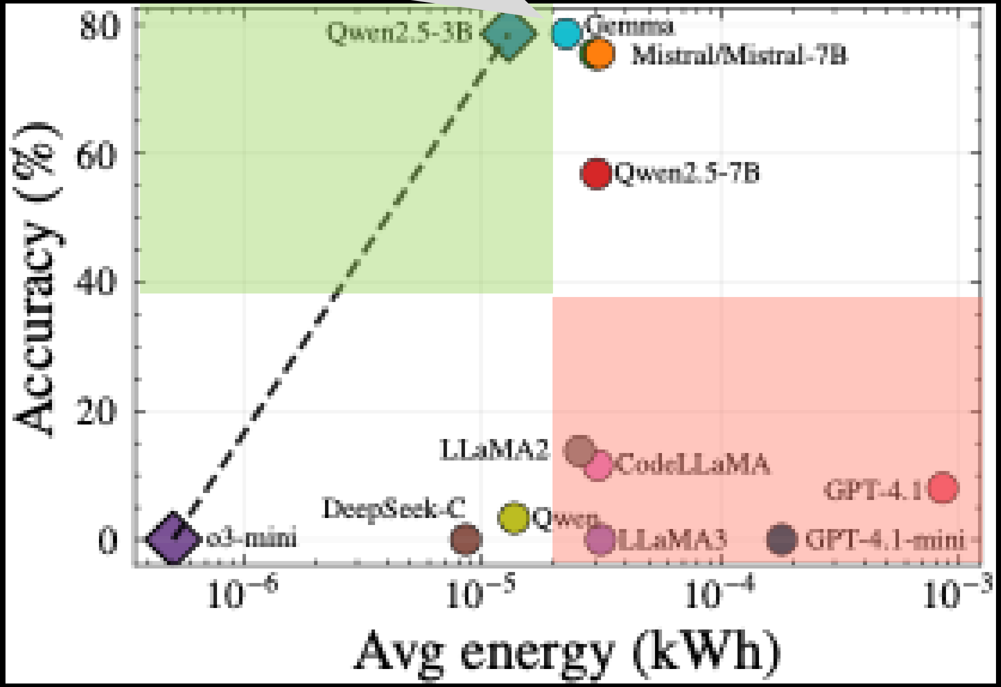
Family 4: Pipeline code generation

EcoLLM Benchmarking Results

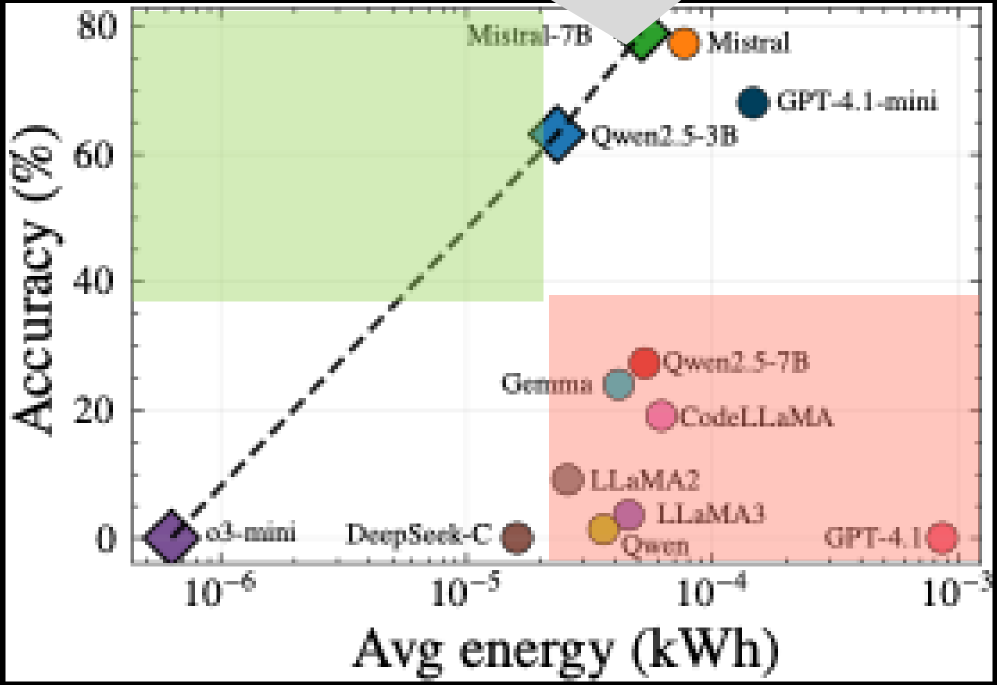
EQ2: Workload + Complexity - How do local and API models behave under increasing task complexity?

Local models dominate simple pipeline generation

Specialized models remain robust as complexity grows



Family 4 and Complexity C1



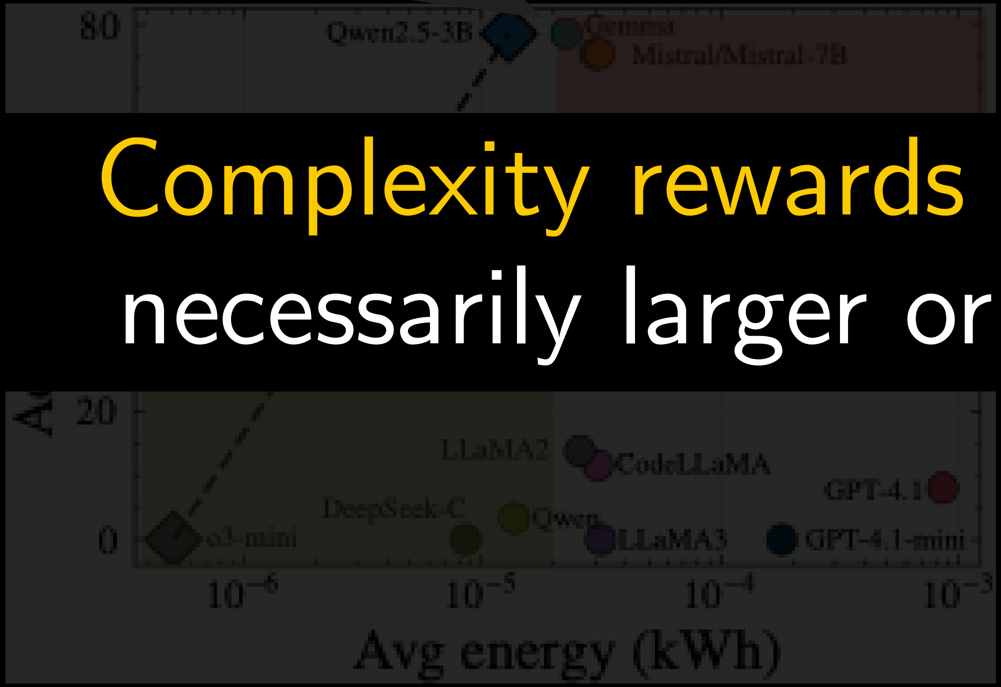
Family 4 and Complexity C4

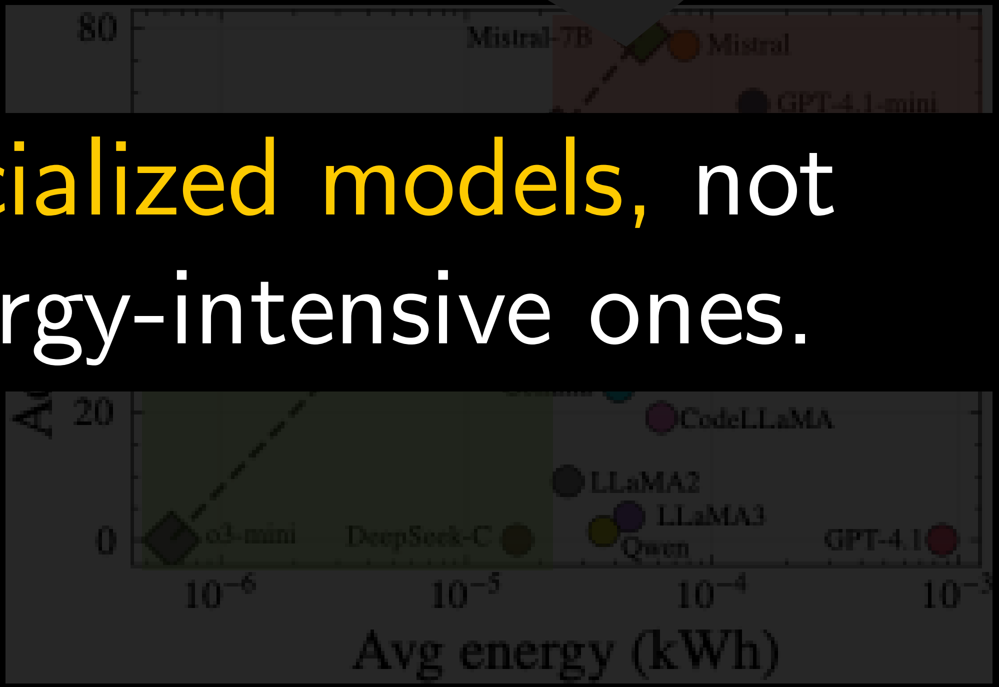
EcoLLM Benchmarking Results

EQ2: Workload + Complexity - How do local and API models behave under increasing task complexity?

Local models dominate simple pipeline generation

Specialized models remain robust as complexity grows





Complexity rewards specialized models, not necessarily larger or energy-intensive ones.

Family 4 and Complexity C1

Family 4 and Complexity C4

Thanks for Your Attention!   Questions

- LLMs are becoming execution-time components in data systems
- How fast? → How scalable? → Monetary cost → **At what energy cost?**
- EcoLLM: energy-aware benchmarking for LLM-based data processing workloads

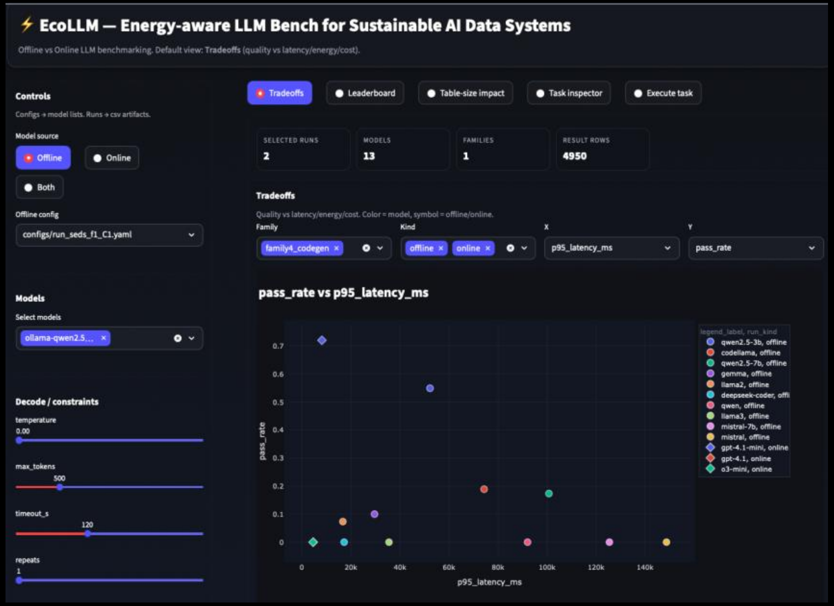
Higher energy consumption does not necessarily lead to better task quality

Model specialization is more important than model size.

Increasing workload complexity **amplifies performance**

Local models can achieve competitive effectiveness at substantially lower energy cost.

Deployment mode strongly shapes energy behavior.



- Energy must become a first-class metric for LLM-based data systems.

Bundesministerium
für Wirtschaft
und Energie

ETAGPT

dfki ai
Deutsches Forschungszentrum
für Künstliche Intelligenz
German Research Center for
Artificial Intelligence

etaGPT – An LLM-based Data Analysis System for Energy Management

etalytics energy intelligence

PTW TU DARMSTADT
Institut für Produktionsmanagement, Technologie und Werkzeugmaschinen

BIONTECH

MERCK

SIEMENS energy

DARZ
Daten sind wertvoller als Geld