



UNIVERSITEIT VAN AMSTERDAM
Instituut voor Informatica

A CASE FOR EXPLICIT KNOWLEDGE MODELS

The Decision Nobody Made

Jan-Christoph Kalo

Which country emitted more CO₂ in 2022?

Australia



United Kingdom



The Decision Nobody Made

Illustrative answer. Values in million tonnes CO₂.

“The United Kingdom emitted more CO₂ than Australia in 2022.”

Australia **384.1**

Territorial accounting

Emissions produced within national borders

United Kingdom **502.8**

Consumption accounting

Emissions attributed to domestic consumption

2022 · Mt CO ₂	Territorial	Consumption
Australia	384.1	365.1
United Kingdom	311.1	502.8

Treating these values as **comparable** is an **unstated decision**.

Fossil and industrial CO₂, excluding land-use change. Fixed GCB / OWID snapshot used in the pilot.

Pilot setup

Question: *Which country had higher total annual CO₂ emissions in 2022?*

We varied the observations and method information in prepared excerpts.

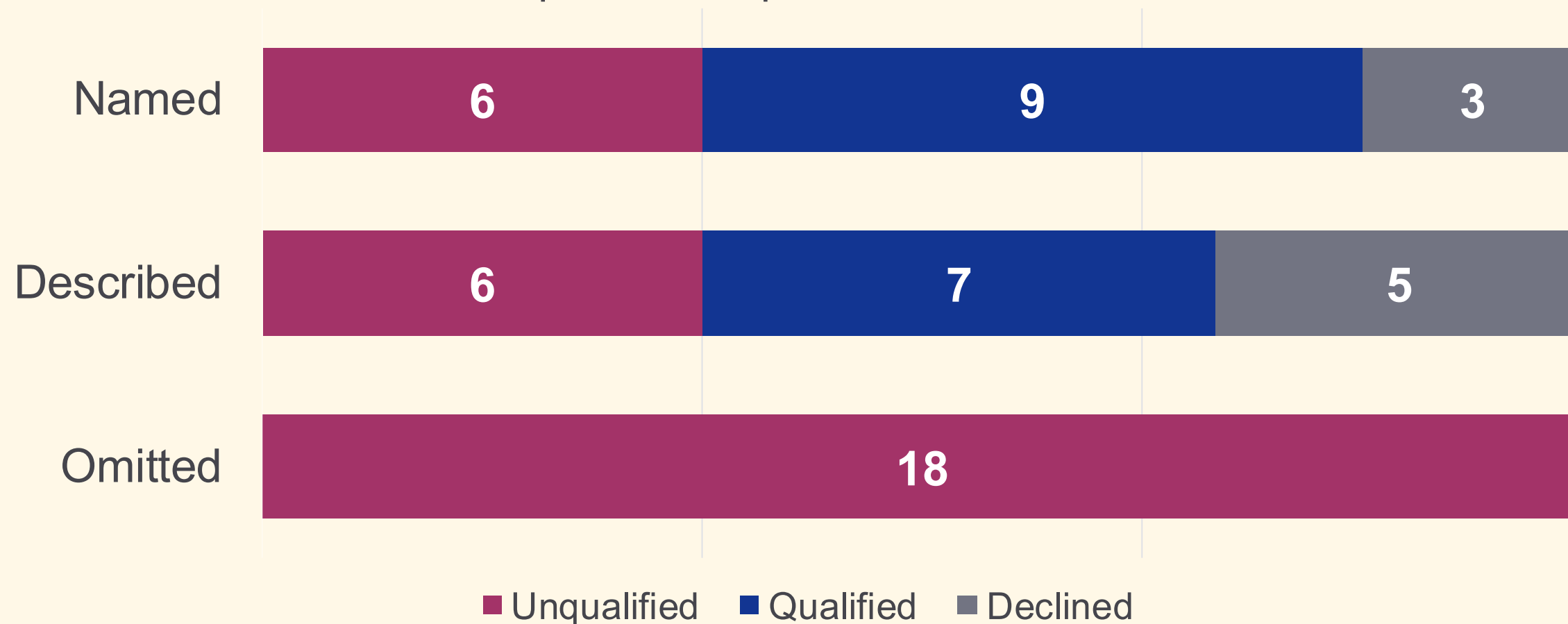
Condition	Information supplied
Mixed: named	Two values using different methods, with names supplied.
Mixed: described	Same values. Methods explained without naming them.
Mixed: omitted	Same values. No method information.
Matched control	Two values under one described method.
Complete control	All four values. Methods named or described.

3 models (Gemma, Opus, Qwen), 3 country pairs, **90 responses**

How did the models handle the comparison?

Mixed methods: 54 responses

Named methods: all 6 unqualified responses were Gemma



Positive controls

18 / 18

Complete labelled input:
gave both rankings

18 / 18

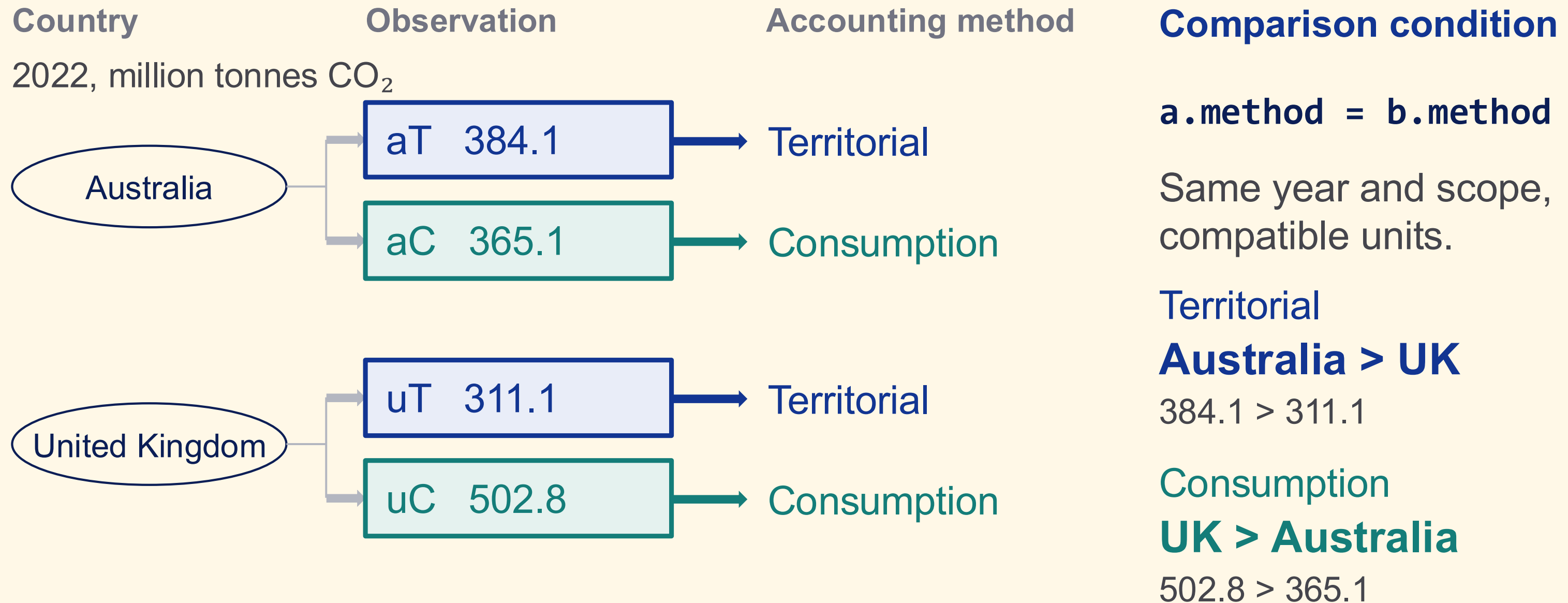
Matched methods:
ranked directly

Unqualified: ranks without a method caveat. Qualified: ranks with one. Declined: no ranking.

All models gave both rankings with **all four observations and their methods**.
 Preparing that input required human **modelling decisions**.

Pilot: 90 responses, three models.

Storing the data in a knowledge graph or database



Explicit modeling **makes it possible to compare comparable values.**
 But what did the user ask for?

Common reasons for explicit representations

Formal reasoning

Derive conclusions using defined rules.

Domain knowledge

Represent constraints and prior knowledge.

Explanations

Inspect assumptions and inference steps.

Learning

Support generalization and data efficiency.

But none of these properties would have prevented our wrong comparison.

Formalisation gives shared knowledge
a **precise computational meaning** to enable communication between agents.

Garcez & Lamb (2020); Xu et al. (2018); Battaglia et al. (2018). Communication: Davis et al. (1993), Gruber (1995).

Do better LLMs solve the problem?

We increasingly **learn the representations**. Suppose the model understands both methods and correctly computes both rankings from all four observations.

Territorial: Australia

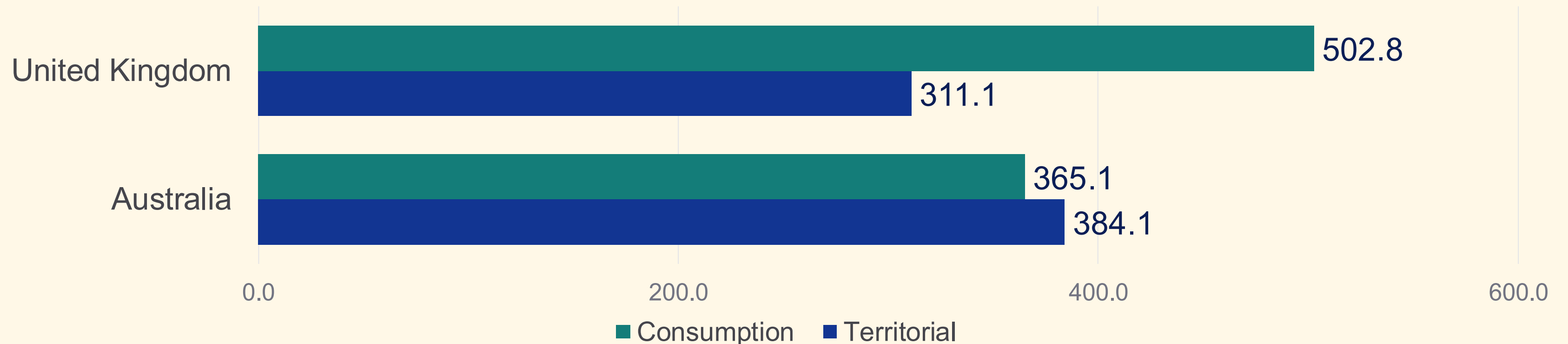
Consumption: United Kingdom

Possible application	What it compares
National emissions inventory	Emissions produced within each country
Consumption footprint	Emissions associated with each country's consumption
Comparison of accounting approaches	Both measures and their different rankings

We should not leave the **meaning of the task** to the **model** that solves it.

Domain assumptions and bias

Do we attribute emissions to production or consumption?
Choosing one of the metrics implies a **human perspective**.



Operationalization is part of the claim.

An LLM may give a default answer.

But explicit models let us examine and keep **different domain perspectives**.

How do we communicate meaning?

Put the definition into the prompt as text.

“Compare using territorial accounting.”

- The prompt names a method.
- We still need to map it to data and define the comparison.

Explicit model and formal specification

The model connects **method definitions** to observations and **comparison rules**.
A table or graph can implement it.

- Select method = Territorial.
- Match year, scope and units.
- Compare the selected values.

Communication

People agree on this specification.
Components refer to the same definitions.

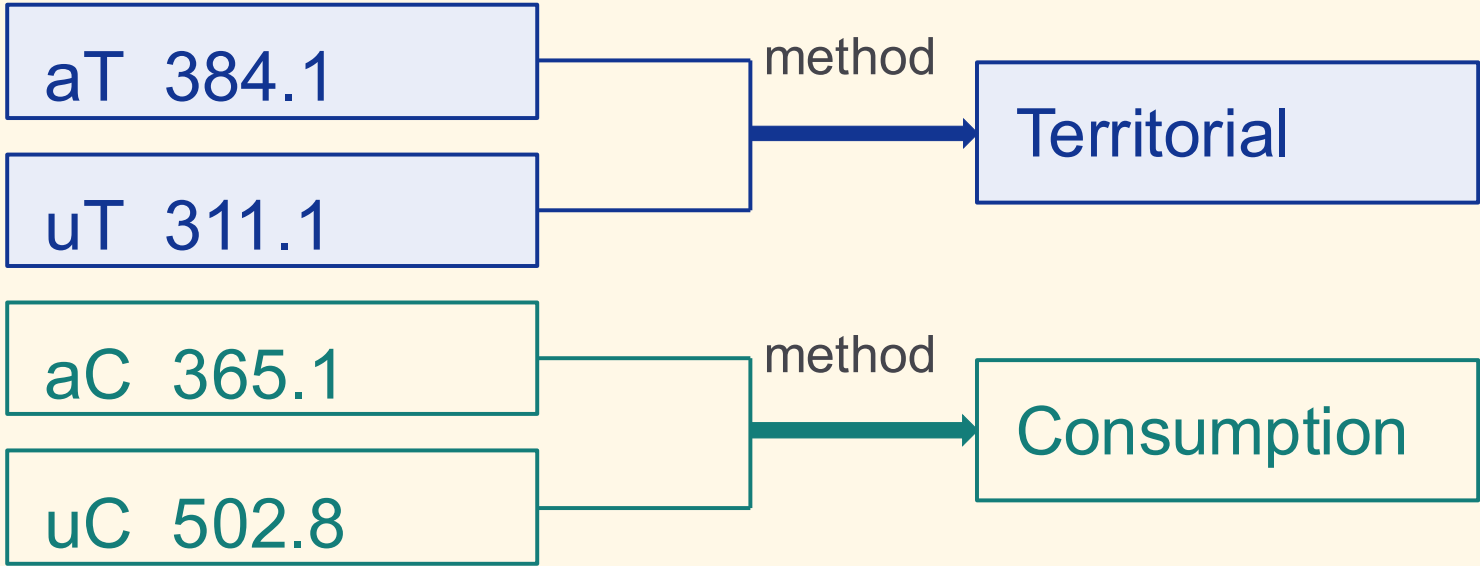
- The query, learned component and evaluator use the same fields and comparison rules.

Explicit models record our decisions about what data means and make those decisions **usable in computation**.

SQL and SPARQL specify the same selection

Method	Territorial
Definition	CO ₂ emitted within national borders.
Source	Our World in Data, CO ₂ dataset metadata

Observation	Value (Mt)	Method
aT	384.1	Territorial
uT	311.1	Territorial
aC	365.1	Consumption
uC	502.8	Consumption



SQL selection

```
SELECT value FROM observations
WHERE method = 'Territorial';
```

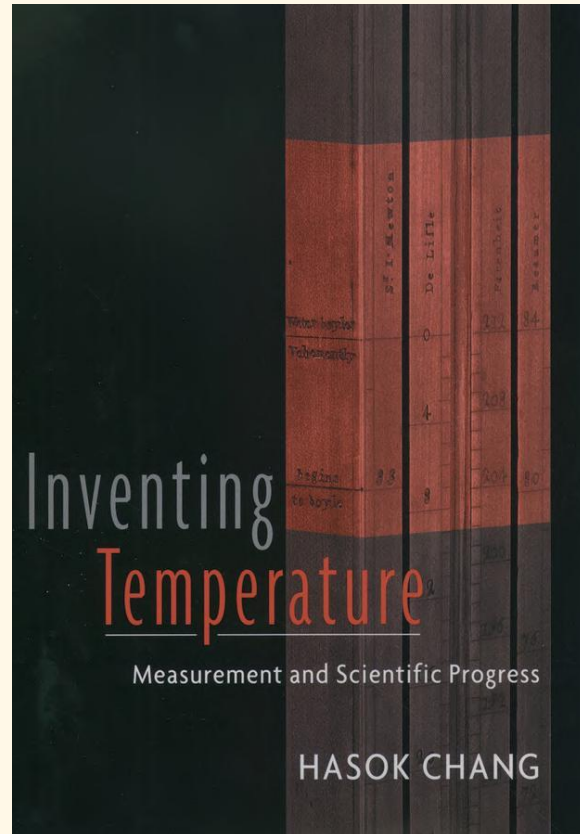
SPARQL pattern

```
?obs :method :Territorial ;
      :value ?value .
```

Both select aT (384.1) and uT (311.1).

The **definition explains** the method. The **queries specify** the selection.

Inventing Temperature



Recommended reading
Hasok Chang (2004)



How can we check a thermometer
before agreeing how to measure
temperature?

Definitions, instruments and standards
develop through **scientific practice**.

We can compare, challenge and revise
conceptual choices.

This is how science moves forward.

Without a definition, the LLM decides at query time what to do with numbers.
An application needs clear definitions.

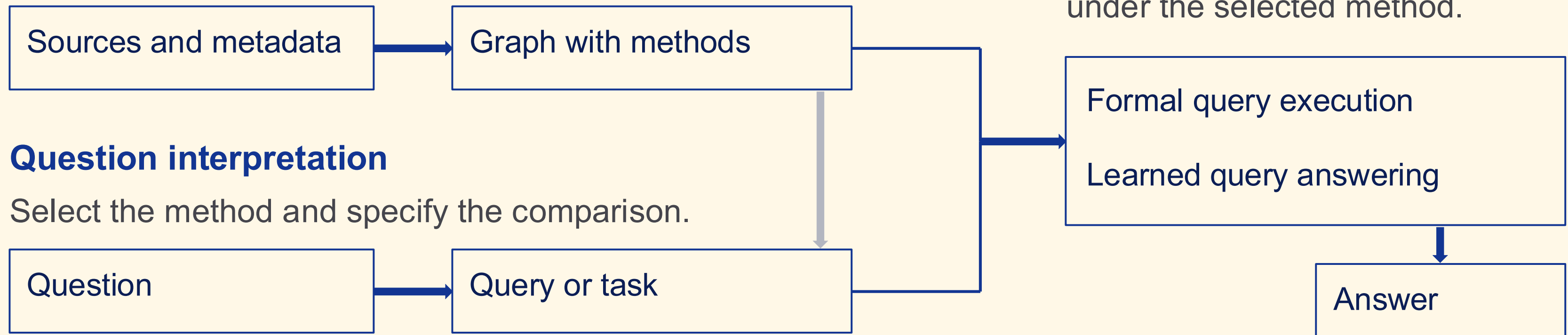
Hasok Chang, *Inventing Temperature* (2004). Instrument: Science Museum Group, 1880.

Explicit definitions guide each part of the system

We retain both accounting methods and define comparisons **within each method**.
That decision specifies what every component must **preserve or compute**.

Construction

Keep the method attached to each observation.



Learning and reasoning

The prediction makes a claim under the selected method.

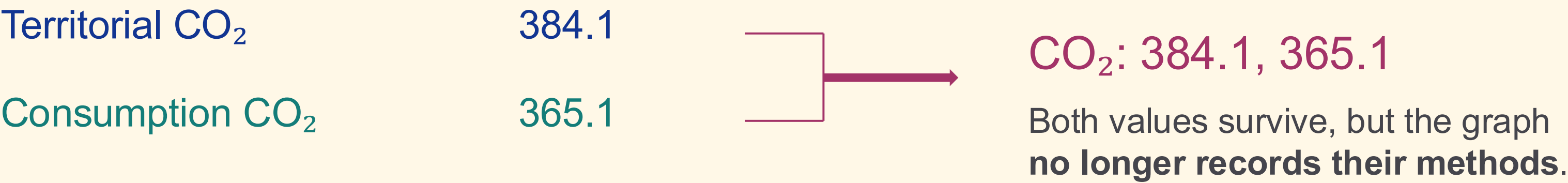
Question interpretation

Select the method and specify the comparison.

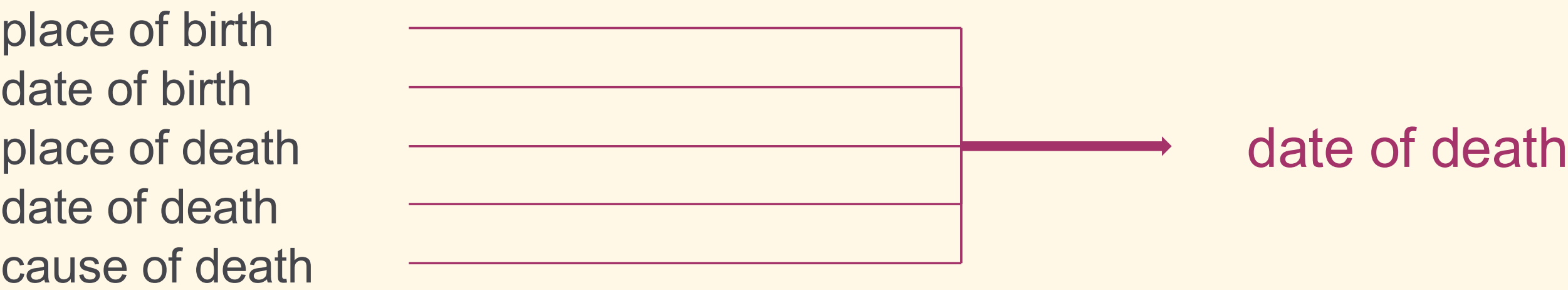
Evaluation uses the **same definitions** to assess construction and answers.

What does a merge assume?

Illustrative extraction: Australia, 2022, Mt CO₂



Reported CESI merge



A merge can **preserve the values** while **losing their meaning**.

CESI example reported in Zhang & Soh, EMNLP 2024, §4.2.2. Emissions extraction is illustrative.

Are automatically generated definitions the solution?

Alan Shepard $\longrightarrow$ participatedIn $\longrightarrow$ Apollo 14

EDC generates a definition and compares it with candidate definitions
“The subject entity took part in the event or mission specified by the object entity.”

Candidate	Definition in the paper (shortened)
mission	Participation in an event or operation
season	Participation in a season of a series
None of the above	No equivalent relation

EDC selects: Alan Shepard $\rightarrow$ mission $\rightarrow$ Apollo 14

Generated definitions might miss what the application needs.
The schema used for extraction determines what any application can do with the data.

Zhang & Soh, Extract, Define, Canonicalize, EMNLP 2024, Figure 1 and §3.1.

A query commits to a measure and an operation

How much CO₂ did the green electricity used in the Netherlands avoid in 2022?

CBS: AvoidedEmission_3

“It covers domestic production of renewable electricity, heat and gas.”

“Imports of green electricity is not included.”

SQL selection

```
SELECT AvoidedEmission_3  
FROM cbs_84918  
WHERE Periods = '2022JJ00';
```

Selecting this field adopts its **reference case** and its restriction to **domestic production**.

A valid query can answer the **wrong question**.
SQL executes the selected measure, including a **mistaken selection**.

LLMs as Knowledge Bases

Completing Wikidata with GPT-3

27 million

estimated additional statements
at 90% precision

Task: Find the developer of
the game **Neko-Nin exHeart**

Context field	Value
Developer	Whirlpool
Publisher	Sekai Project

GPT-3: Whirlpool, Sekai Project

Correct: Whirlpool

GPT-3 knows both facts, but answers as if developer
meant any company behind the game.

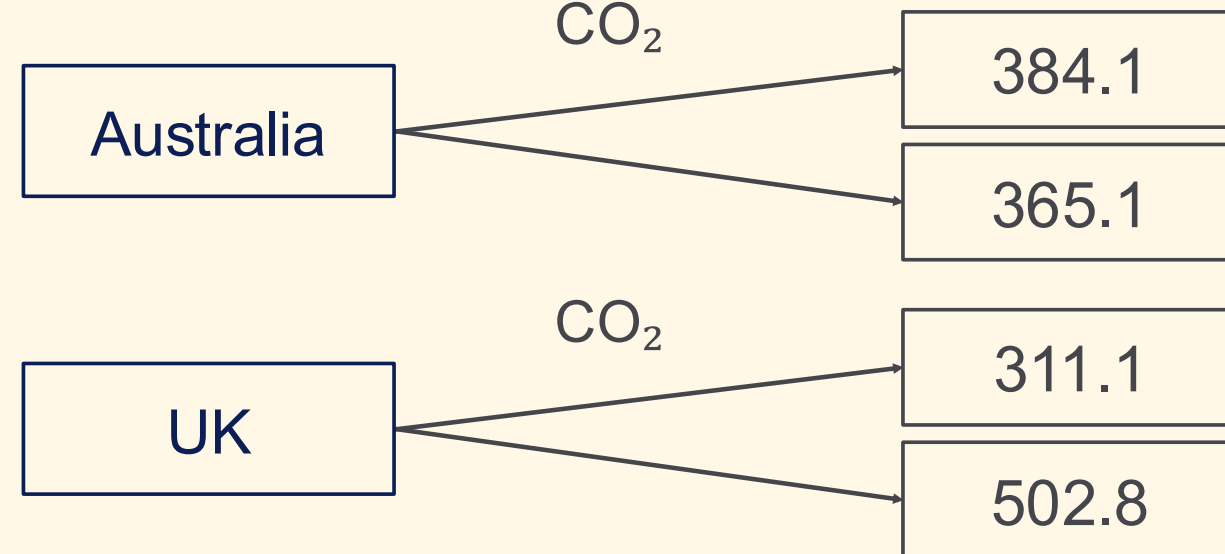
Without the defined relation, a wrong fact is used in the application.

Veseli, Razniewski, Kalo & Weikum, Findings EMNLP 2023. Table 2 and Appendix A.2.

What does a prediction over this graph mean?

Country and value

Same values: 2022, Mt CO₂

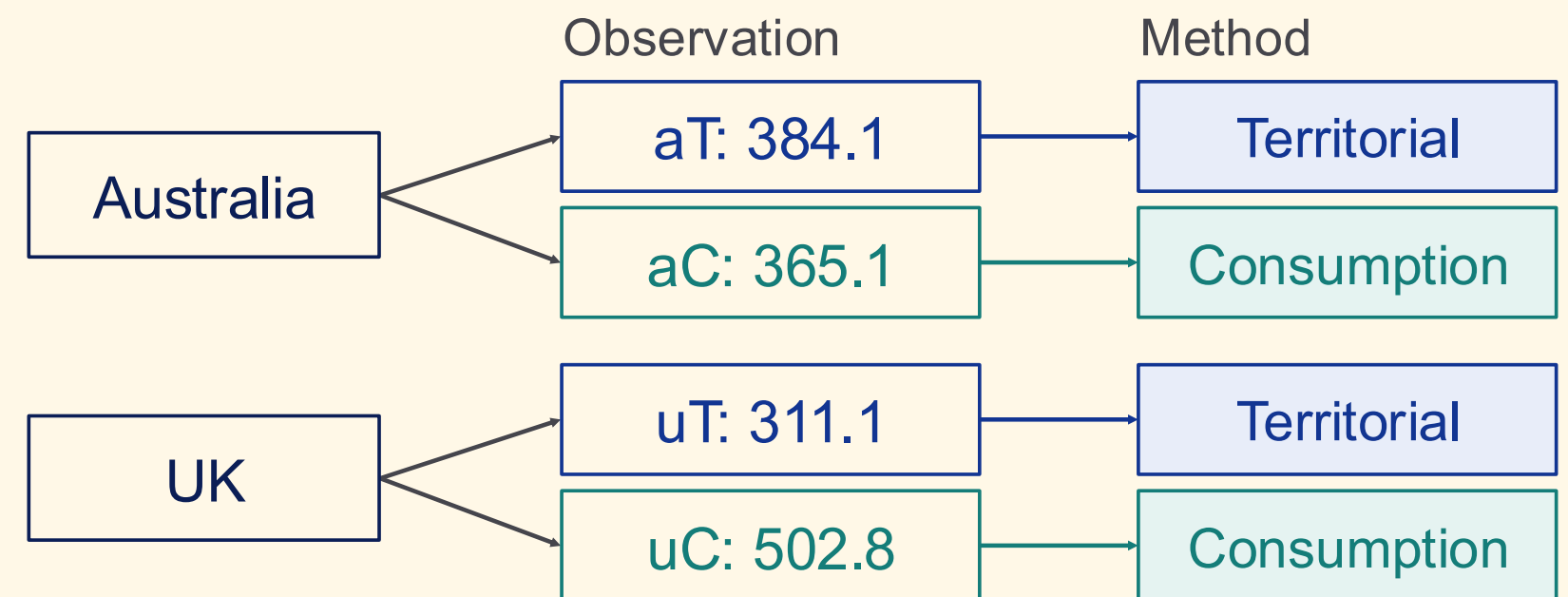


No value records its method,
so a ranking can mix them.

Each ranking is a claim under one method:

The graph decides what a prediction claims.

Observations with accounting methods



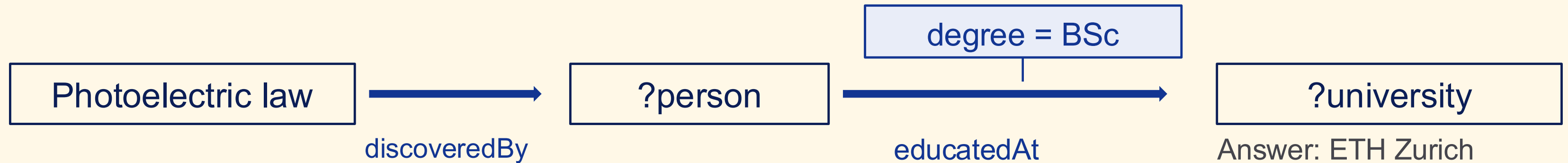
Every observation records its method.

Territorial: Australia 384.1 > UK 311.1

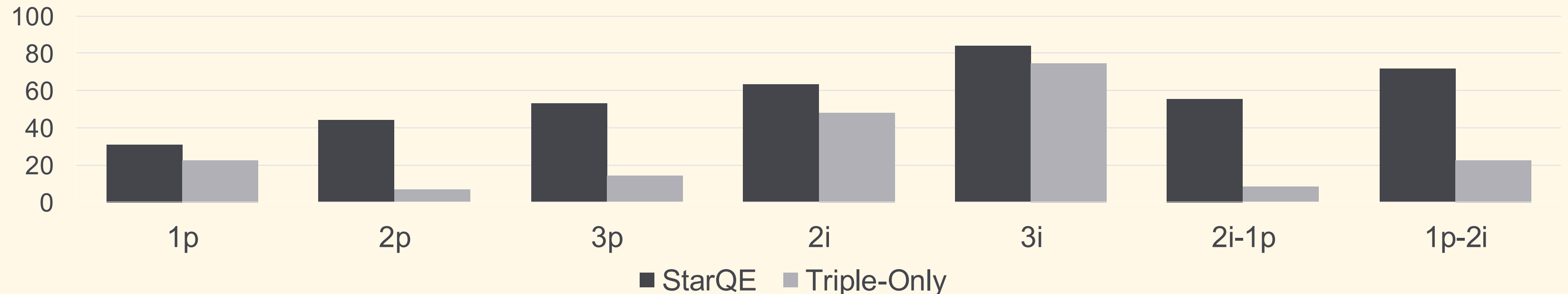
Consumption: UK 502.8 > Australia 365.1

Explicit qualifiers improve learned query answering

Where did the discoverer of the photoelectric law obtain a BSc?



MRR (%) with and without qualifiers



StarQE scores higher on all seven query patterns.

Without the qualifier, the model answers a broader question: any degree.
The qualification **stays explicit** while the computation **is learned**.

Scoring the final answer

System answer

Country-only reference answer

United Kingdom

United Kingdom

The observations behind the answer

Country	Value (Mt CO ₂)	Method
United Kingdom	502.8	Consumption
France	295.3	Territorial

A correct answer can be based on an **unsupported comparison**.
Check the **observations and comparison conditions**.

Recorded pilot responses with methods omitted. Both valid UK–France comparisons also return UK.

The Decision Nobody Made?

Which country emitted more CO₂ in 2022: Australia or the UK?

Illustrative scoring example. No accounting method is specified.

Reference answer

Australia

Territorial: 384.1 > 311.1

System answer

United Kingdom

Consumption: 502.8 > 365.1

Exact-match score: **0 / 1**

The reference answer **adopts territorial accounting**.

A reference answer can turn an **unstated interpretation** into **ground truth**.

Explicit domain models for AI systems

Construction

The extraction schema decides which distinctions survive.
Preserve what later steps must tell apart.

Question interpretation

The selected field decides which question gets answered.
Use the metadata to make that choice explicit.

Learning and reasoning

The graph decides what a prediction claims.
The representation behind it can still be learned.

Evaluation

The gold answer decides which interpretation counts as correct.
Check the observation behind the answer, not just the answer.
A different definition should select different observations.

A new model should **improve the computation**
without **silently changing the task**.



The Decision Nobody Made

Explicit models record human **decisions about what data means** and make those decisions **usable in computation.**

People and systems can share definitions while using different internal representations.

We should be able to improve the model without leaving the **meaning of the task** for the model to decide.