Operationalization Is Part of the Claim

Operationalization-Aware Knowledge in the Era of Generated Answers

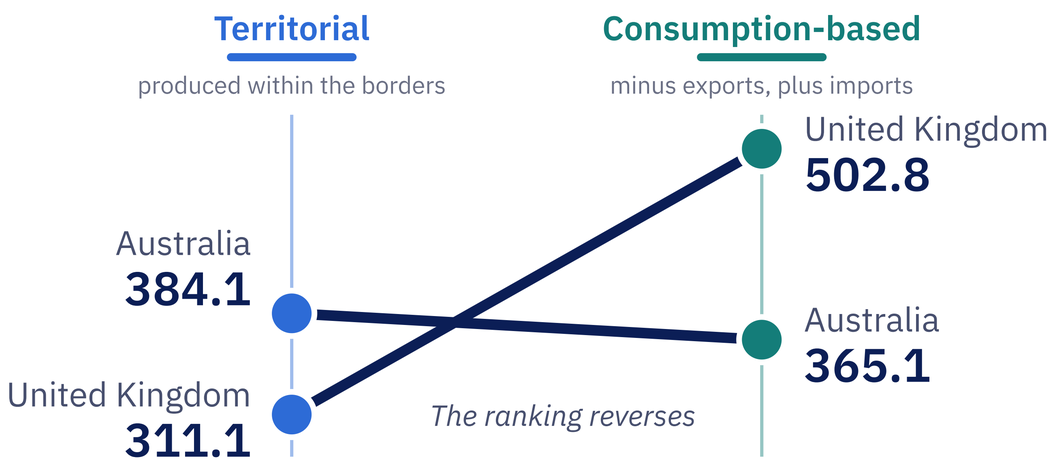
Jan-Christoph Kalo j.c.kalo@uva.nl

INDE lab
University of Amsterdam

1 Operationalization plurality

Did the United Kingdom or Australia emit more CO_2 in 2022?

Operationalization: how a quantity is defined and how it is estimated.



$MtCO_2$ in 2022, fossil fuels and industry. Global Carbon Budget via Our World in Data (2025).

Both are correct once their operationalizations are made explicit.

An adequate answer

```
Australia under territorial accounting,
the UK under consumption-based accounting.
The question does not specify which.
```

Factual conflict

Same quantity and operationalization
One value may be wrong
Data quality helps decide which is right

Work on knowledge conflicts:
Xie et al. 2024; Cattan et al. 2025

Operationalization plurality

Differently defined quantities
Neither value need be wrong
Data quality cannot decide which is relevant

Also in unemployment, inflation, excess mortality and food insecurity

2 Operationalization collapse

Retrieval may return one value per country, each under a different framework.

Question: `Did the United Kingdom or France have higher total annual CO2 emissions in 2022?`

Frameworks omitted from model input

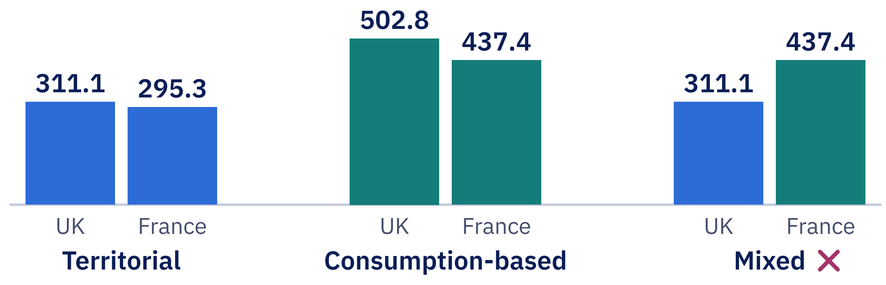
Retrieved
United Kingdom — **311.1 $MtCO_2$**
● territorial

Retrieved
France — **437.4 $MtCO_2$**
● consumption-based

Framework labels added for the reader.

Gemma 4 31B IT → **France** ✗
Qwen3.8 Max → **France** ✗
Claude Opus 5 → **France** ✗

The UK is higher under both frameworks.



Operationalization collapse: presenting one result as if it did not depend on the operationalization, or making unsupported comparisons across operationalizations.

Controls: with both values from the same framework, all 18 ranked directly. With all four values, all 18 presented both rankings.

The pilot shows that the problem exists, not how often it occurs.

Pilot: 54 mismatched requests using constructed passages; one square per response



Six responses per cell: three country pairs × two framework assignments.

3 What changes with generated answers

Would a benchmark notice?

Germany or Japan?
Japan ✓ matches the gold answer

In the pilot, all unqualified rankings of Germany and Japan selected Japan, which is higher under both frameworks. A benchmark that scored only the selected country would have marked them correct.

What should a ranked answer say?

1 **Japan** 2 **Germany**

Territorial: 3 **Australia** 4 **UK** 5 **France**
Consumption-based: 3 **UK** 4 **France** 5 **Australia**

Adding irrelevant qualifications, or refusing comparisons that the claims support, is also wrong.

Who decides which values can be compared?

Previously
Data producers
Domain standards
Human analysts
→ **Now**
One generated answer
drawing on an open set of sources

If these decisions remain hidden, operationalization choices appear as neutral facts.

This is especially important in automated meta-analysis, where results must be comparable before they are combined. Qiu et al. 2025; Xie et al. 2026

4 Open questions

CHALLENGE 1: RECOVERY

Which framework is behind 311.1 Mt?

United Kingdom — CO_2 emissions → **311.1 $MtCO_2$**
↳ point in time: 2022
↳ stated in: Our World in Data
↳ determination method (P459): ?

Existing models: Wikidata qualifiers, RDF Data Cube, GSIM, SOSA.

The definition is often documented elsewhere, so retrieval may find the value but miss what it means. Attaching the wrong operationalization can be more misleading than leaving it unresolved.

CHALLENGE 2: COMPARISON

Can 311.1 and 437.4 appear in one answer?

● UK 311.1 ● France 437.4

✓ Report both, each with its framework
✗ Rank directly from these two values alone

It depends on the requested comparison or combination. Explicit mappings or justified bounds can support it; otherwise it stays unresolved.

CHALLENGE 3: SELECTION

Australia or the United Kingdom?

Metadata shows the valid answers; the basis for choosing must come from the question, its context, or a convention.

- ● Emissions inventory → **Australia**
- ● Consumption footprint → **UK**
- ● Reporting convention → **Australia** (chosen by Qwen and Claude)
- ○ Not stated → state the assumption, show both, or ask

A convention may provide a basis within its intended context, but not for every purpose.

How this differs from related work

Ambiguity Min et al. 2020
Ambiguity work makes interpretations explicit; we focus on which comparisons and combinations the claims support.

Data integration
Goh et al. 1999; Lenzerini 2002; Kroll et al. 2020
It usually relies on schemas or mappings specified in advance; generated answers must decide at answer time which comparisons the claims support.

Source selection Zhang et al. 2026
It finds a sufficient set of sources, but that set can contain two valid, differently defined values.

A value may be correct while the answer is still wrong.

Our vision: keep operationalizations attached to claims, and choices among them explicit, until the question, its context, or an established reporting convention provides a basis for resolving them.

Code, prompts, responses and annotations
`github.com/JanKalo/akbc2026-vision`