

Operationalization Is Part of the Claim: Operationalization-Aware Knowledge in the Era of Generated Answers

Jan-Christoph Kalo
University of Amsterdam
j.c.kalo@uva.nl

Abstract

LLM-based systems increasingly search across sources and synthesize numerical findings, but can treat values from different sources as directly comparable. The same phenomenon may be measured by defining the quantity differently or by estimating it through different procedures. Together, these choices form a value’s *operationalization*. Values produced under different operationalizations can differ without creating a factual conflict. Ignoring such differences can lead systems to make unsupported comparisons or combinations, or to present one result as the single correct answer even though the appropriate choice depends on the user’s purpose. We call this *operationalization collapse*. This vision paper argues that operationalization is part of a numerical claim, not optional metadata. Operationalization information can show what a value means and which comparisons or combinations it supports, but it may not determine which valid result should be used to answer the user’s question. We outline a research program for recovering operationalizations, attaching them to claims, determining which comparisons and combinations those claims support, and generating answers that preserve unresolved choices.

1 Introduction

Instead of reading several sources and synthesizing the findings manually, users increasingly turn to generative search systems for complex tasks (Suri et al., 2024; Aral et al., 2026).

Consider a seemingly simple question: *Did the United Kingdom or Australia have higher total annual CO₂ emissions in 2022?* An AI system searching for the relevant facts might encounter the four values shown in Table 1. National CO₂ emissions are reported using several accounting frameworks that define the measured quantity differently. Territorial emissions include emissions

Accounting framework	Australia	United Kingdom
Territorial	384.1	311.1
Consumption-based	365.1	502.8

Table 1: Total annual CO₂ emissions from fossil fuels and industry, excluding land-use change, in 2022, in MtCO₂. Data: [Our World in Data \(2025\)](#)

produced within a country’s borders. Consumption-based emissions subtract the emissions released in producing exported goods and add the emissions released abroad in producing imported goods. The two frameworks assign emissions to countries differently.

The answer depends on which operationalization is used. Australia had higher territorial emissions, whereas the UK had higher consumption-based emissions. These claims do not contradict each other: both are correct once their operationalizations are made explicit. We call the coexistence of multiple valid operationalizations for the same phenomenon *operationalization plurality*.

Recent work on knowledge conflicts in LLMs studies how models respond when retrieved sources disagree or when retrieved evidence conflicts with parametric knowledge (Xie et al., 2024; Kortukov et al., 2024; Xu et al., 2024; Cattani et al., 2025). Our example differs: the values refer to differently defined quantities and only appear contradictory when their operationalizations are omitted.

An AI system should not silently erase this difference. Depending on the question and the user’s purpose, it could ask for clarification, state which operationalization it uses, or present both answers and explain why they differ. If it ignores the distinction, three failures are possible. It may select one operationalization and present the result as if it did not depend on that choice. It may compare values produced under different operationalizations and draw an invalid conclusion. Or it may treat the differing values as estimates of the same quantity

and conclude that one of them must be wrong. The first two silently remove an unresolved operationalization difference that changes the answer. We call this *operationalization collapse*.

An adequate answer can preserve this unresolved choice without becoming lengthy:

Australia had higher territorial CO₂ emissions in 2022: 384.1 versus 311.1 MtCO₂. Under consumption-based accounting, however, the United Kingdom was higher: 502.8 versus 365.1 MtCO₂. The question does not specify which accounting framework to use.

This answer does not select one operationalization as the single correct one, and it does not combine values from different operationalizations.

A small pilot study illustrates the difficulty of maintaining this distinction (Section 3.1). We gave three LLMs constructed passages reporting countries' emissions under different accounting frameworks. When the passages omitted the frameworks, all three models ranked the countries without qualification. Unqualified rankings also occurred when the frameworks were named or described, suggesting that making operationalizations explicit is not always sufficient for models to respect them when comparing values.

The problem is not unique to LLMs. Database and Semantic Web research has long addressed semantic heterogeneity through schemas, mappings, ontologies, and context models (Lenzerini, 2002; Zheng et al., 2012; Kroll et al., 2020). Systems that generate answers may instead draw on an open set of sources and must decide at answer time which comparisons or combinations the claims support. This is especially important in automated meta-analysis, where systems retrieve and combine findings across many studies (Qiu et al., 2025; Xie et al., 2026).

Even complete and correct operationalization information would not always determine the answer. Territorial accounting may be appropriate for a national emissions inventory, whereas consumption-based accounting may be appropriate for assessing a country's consumption footprint. If the question does not indicate which purpose matters, and no established convention resolves the choice, selecting one result would introduce an additional assumption. A system should make that assumption explicit, present both answers, or ask for clarification.

Our vision is to move from treating numerical values as standalone facts to treating their operationalizations as part of the claim, so that generated

answers preserve plurality without obscuring when one answer is justified.

2 Related Work

Knowledge conflicts and question ambiguity. Work on knowledge conflicts in retrieval-augmented generation examines disagreements between retrieved sources and between retrieved evidence and a model's stored knowledge (Xie et al., 2024; Kortukov et al., 2024; Xu et al., 2024). This literature also recognizes that different answers need not conflict. Cattan et al. (2025) distinguish complementary information from factual conflict and study responses that preserve relevant alternatives. Similarly, work on question ambiguity identifies multiple plausible answers and makes their interpretations explicit, including numerical answers that depend on inclusion criteria (Min et al., 2020). We focus on how operationalization differences constrain the use of numerical claims together. Values may all be valid under their respective quantity definitions while failing to support a direct comparison or combination. Making those definitions explicit can clarify which operations are supported, but may still leave unresolved which quantity is relevant to the user's question.

Data integration. Data integration research has long addressed differences between heterogeneous sources through schemas, mappings, and explicit descriptions of context (Lenzerini, 2002; Rahm and Bernstein, 2001; Goh et al., 1999; Zheng et al., 2012; Kroll et al., 2020). These approaches identify how data from different sources correspond and under which conditions they can be combined. They usually rely on schemas, mappings, or integration goals specified in advance. Systems that generate answers may instead draw on claims from sources without a common schema or mapping. For each question, they must decide at answer time which comparisons or combinations those claims support.

Statistical metadata. Standards and systems such as the Generic Statistical Information Model, the RDF Data Cube vocabulary, and Data Commons can represent numerical observations together with information about what they measure, how they were produced, and where they come from (United Nations Economic Commission for Europe, 2024; Cyganiak and Reynolds, 2014; Guha et al., 2023). Recent work also makes this information available to language models for statistical

retrieval, reasoning, and source selection (Radhakrishnan et al., 2024; Belyaeva et al., 2026; Zhang et al., 2026). In particular, Zhang et al. (2026) use metadata to identify relevant and sufficient data sources for an analytical task. We focus on the subsequent problem: when several retrieved claims remain valid under different operationalizations, the system must determine which comparisons, combinations, and answers they support. Metadata alone may still not determine which claim should answer the question when the user’s purpose is unclear.

Automated evidence synthesis. LLM systems increasingly support systematic reviews and meta-analyses by retrieving studies, screening them, and extracting findings for synthesis (Qiu et al., 2025; Xie et al., 2026; Tan and D’Souza, 2026). This makes our problem especially important: before results can be combined, the system must determine whether they are comparable given how the quantities were defined and estimated. Existing work mainly studies retrieval, screening, and extraction. It generally does not address which comparisons or combinations are valid across such findings, or how unresolved differences should be preserved in the final synthesis.

3 Operationalization Plurality Does Not Imply Factual Conflict

A construct is the concept or phenomenon we want to measure, such as unemployment, emissions, or food insecurity. Operationalizing a construct requires two kinds of decisions. First, we must define the quantity to be measured: what is included or excluded, how observations are counted, and how values are aggregated (Adcock and Collier, 2001; Jacobs and Wallach, 2021). Second, we must decide how to estimate that quantity, including the data sources, instruments, models, estimators, and adjustments used. Together, the definition of the quantity and the procedure used to estimate it form an operationalization. We focus on numerical claims because operationalization differences directly affect comparisons, combinations, and rankings, although the same problem can also arise for other factual claims.

Different operationalizations may define different quantities for the same construct. They may also use different procedures to estimate the same quantity. In both cases, the resulting values may differ even though neither claim is necessarily wrong. This is clearest when the operationalizations define

different quantities. When different procedures estimate the same quantity, the boundary with factual conflict is less clear, and credibility, uncertainty, and data quality may matter. The research program in Section 4 focuses on differently defined quantities. For factual conflicts between estimates of the same quantity, the methods discussed in Section 2 apply.

In our example, territorial and consumption-based accounting define different quantities rather than competing estimates of the same quantity. Similar operationalization differences also arise in measures of unemployment, inflation, excess mortality, and food insecurity.

A clear factual conflict, by contrast, arises when two claims report incompatible values for the same quantity, entity, and period under the same operationalization. Two sources might, for example, report different values for Australia’s territorial emissions in 2022 while using the same accounting framework, data version, and estimation procedure. The difference might then result from an error in calculation or reporting. Credibility, uncertainty, and data quality become relevant. Territorial and consumption-based emissions instead describe different quantities. Their values therefore do not conflict merely because they differ.

Different operationalizations do not always converge into a single standard. In some fields, one operationalization becomes widely accepted (Chang, 2004). In others, several operationalizations persist because they define quantities that serve different legitimate purposes. Better data or estimation procedures may improve the estimates within each operationalization, but they do not determine which quantity is relevant to the question.

Whether claims can be used together depends on the requested comparison or combination. Credibility, uncertainty, and data quality may help choose among estimates of the same quantity, but they cannot determine which of two differently defined quantities is relevant.

3.1 Pilot study

We examined how models compare countries when the retrieved sources report emissions under different accounting frameworks. Each passage was about 200 words long and reported one country’s emissions in the style of a data release or a policy brief. We varied a single sentence to name the framework, describe it without naming it, or provide no operationalization informa-

Passages	Ranked without qualification	Ranked with qualification	Declined
Framework named	6	9	3
Framework described	6	7	5
No operationalization	18	0	0

Table 2: Mismatched responses by operationalization information, with 18 responses per row.

tion. Mismatched requests contained one passage per country under different frameworks and asked which country had higher total annual CO₂ emissions in 2022, similar to the introduction’s example. We used three country pairs: UK–France, Australia–UK, and Germany–Japan and three models: Gemma 4 31B IT, Qwen3.8 Max, and Claude Opus 5 (Appendix A).

Two kinds of control requests used the same passages. Matched requests contained two passages under the same framework, so a direct ranking was supported. Complete requests contained all four values, as in the introduction, so an adequate answer could preserve both rankings. All 18 matched responses ranked directly. All 18 complete responses presented both rankings. One Gemma response first paired values across frameworks before correcting itself. For Australia and the United Kingdom, Qwen and Opus then selected the territorial result as the one conventionally reported.

Table 2 reports 54 mismatched responses. The two values alone do not justify a ranking on a common accounting basis, since they use different operationalizations. We count rankings without qualification, rankings qualified by framework differences, and whether the LLM declines to rank both countries. The validity of the qualification of the model is not further investigated. In no response the LLMs asserted an error or substituted the provided values its own values.

When the passages did not identify the operationalizations, all 18 responses ranked them without qualification. For example, when the passages gave the United Kingdom’s territorial value (311.1) and France’s consumption-based value (437.4), all three models answered France, which is wrong under both frameworks.

With the framework named or described, Gemma ranked without qualification in all 12 responses. Qwen ranked with qualification in 11 and declined in 1. Opus declined in 7 of the 8 responses for UK–France and Australia–UK. For Germany–Japan, it ranked with qualification. Naming and de-

scribing the frameworks produced similar response patterns. Opus, and sometimes Qwen, invoked trade adjustments not documented in the passages.

Providing operationalization information did not consistently prevent unqualified comparisons. The pilot does not test recovery of operationalizations documented separately from values.

4 From Operationalization Plurality to an Answer

Addressing operationalization plurality in generated answers raises three connected research challenges. They concern recovering and retaining the operationalization of each value, determining which comparisons and combinations the resulting claims support, and deciding which supported result adequately answers the question.

The first challenge is to recover the operationalization of each value and keep it attached to the correct claim. Rather than treating a value as a standalone fact, the system must represent it as part of a claim that specifies the construct and quantity measured, the entity (here, a country) and time period, the unit, the operationalization, and the provenance. For example, the first value in Table 1 connects 384.1 MtCO₂ to Australia in 2022 to emissions from fossil fuels and industry excluding land-use change, to territorial accounting, and to its source. Existing models for statistical observations, variables, provenance, and qualified statements, including Wikidata’s determination-method qualifier (P459), provide useful building blocks for representing such claims without requiring a universal ontology of operationalizations (United Nations Economic Commission for Europe, 2024; Cyganiak and Reynolds, 2014; Janowicz et al., 2019; Vrandečić and Krötzsch, 2014).

Recovering the operationalization remains difficult even where suitable metadata models exist (United Nations Economic Commission for Europe, 2024; Cyganiak and Reynolds, 2014; Radhakrishnan et al., 2024; Belyaeva et al., 2026). A value and its operationalization may be documented in different places, and the documentation may not mention the entity in question. Retrieval based mainly on topical relevance may therefore find the value while missing the information needed to interpret it. The system must determine which description applies to which value and indicate when information about the operationalization remains incomplete or uncertain. Attaching the wrong operationalization

to a value can be more misleading than leaving it unresolved.

The second challenge is to assess whether the claims support the requested comparison or combination. In the example, the territorial values for Australia and the UK support one ranking, while the consumption-based values support another. Retrieval may return Australia's territorial value from one source and the UK's consumption-based value from another. Directly ranking the countries from those two values alone is unsupported because they refer to differently defined quantities. The four claims in Table 1 can nevertheless be presented together to explain why the rankings differ. Two values are therefore not simply compatible or incompatible. Whether their claims can be used together depends on the requested comparison or combination. At answer time, the system must determine whether those observations support the requested operation. In the pilot, unqualified comparisons persisted even when the passages identified the operationalizations. Explicit mappings or justified bounds may support comparisons across operationalizations. Otherwise, this judgment may require inference from descriptions of the operationalizations (Lenzerini, 2002; Goh et al., 1999; Zheng et al., 2012; Kroll et al., 2020). When this cannot be established, the requested comparison or combination should remain unresolved.

The third challenge is to determine which of the supported results adequately answers the question. Operationalization information can identify the valid alternatives and explain their differences, but it does not necessarily provide a basis for selecting among them. That basis must come from the question, its context, or an established reporting convention. Such a convention may provide a basis within its intended context, but it does not make that operationalization appropriate for every purpose. An answer that relies on a reporting convention should state which operationalization it uses, as Qwen and Opus did when all four values were available. If the question, its context, and reporting conventions do not resolve the choice, the system should preserve the alternatives rather than silently selecting one result (Min et al., 2020; Cattani et al., 2025; Zhang et al., 2026).

A good system therefore preserves operationalization plurality when the choice of operationalization changes the answer, and neither the question, its context, nor an established reporting convention provides a defensible basis for selecting one result.

Otherwise, it may answer directly while retaining the relevant qualification. Silently removing an unresolved choice that changes the answer produces operationalization collapse.

Evaluation must reflect both sides of this requirement. Systems may fail by collapsing operationalization differences, but also by surfacing irrelevant qualifications or refusing comparisons that the claims support. Evaluation therefore needs cases in which the choice of operationalization changes the answer and cases in which a direct answer is justified. A benchmark with a single gold answer may itself encode one operationalization while presenting the target as independent of that choice. Matching that value alone cannot distinguish a system that correctly operationalizes it from one that silently collapses the difference or merely expresses general caution. In the pilot, all unqualified rankings for Germany and Japan selected Japan, which has higher emissions under both frameworks. A benchmark that scored only the selected country would have marked these responses correct.

5 Conclusion and Outlook

Generated answers concentrate decisions that were previously distributed across data producers, domain standards, and human analysts. When systems draw on values from an open set of sources, they implicitly decide which values can be compared or combined and which operationalization should be used to answer the question. If these decisions remain hidden, operationalization choices appear as neutral facts.

Our vision is to 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. The consequence is that the correctness of an answer cannot be judged from the selected value alone: a value may be correct while the answer is still wrong because it hides a choice that changes the result. Future work must determine when such choices matter in practice and whether systems can preserve them without treating every difference in operationalization as an obstacle to answering.

Limitations

This paper presents a conceptual framework rather than a formal semantics or an implemented system. Operationalizations may be only partly documented, and judgments about whether claims sup-

port a particular comparison or combination may vary across domains and reporting contexts. The question, its context, or an established reporting convention may also fail to provide a clear basis for selecting among valid claims. Finally, the prevalence and practical impact of operationalization collapse in generated answers remain empirical questions. The pilot uses passages we wrote for one accounting distinction, one data source, and three country pairs, with one response per model and request. Thus we use it to demonstrate the existence of the problem itself, without making statements about the prevalence of the error itself.

References

- Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546.
- Sinan Aral, Haiwen Li, and Rui Zuo. 2026. [The rise of AI search: Implications for information markets and human judgement at scale](#). *arXiv preprint arXiv:2602.13415*.
- Irina Belyaeva, Christopher Carino, and Liang-Chi Wang. 2026. [Leveraging survey metadata for LLM reasoning via knowledge graphs](#). *Journal of Data Science*, pages 1–22. Advance online publication.
- Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roei Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. 2025. [DRAGged into conflicts: Detecting and addressing conflicting sources in search-augmented LLMs](#). *arXiv preprint arXiv:2506.08500*.
- Hasok Chang. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.
- Richard Cyganiak and Dave Reynolds. 2014. [The RDF data cube vocabulary](#). W3C recommendation, World Wide Web Consortium.
- Cheng Hian Goh, Stéphane Bressan, Stuart E. Madnick, and Michael D. Siegel. 1999. [Context interchange: New features and formalisms for the intelligent integration of information](#). *ACM Transactions on Information Systems*, 17(3):270–293.
- Ramanathan V. Guha, Prashanth Radhakrishnan, Bo Xu, Wei Sun, Carolyn Au, Ajai Tirumali, Muhammad J. Amjad, Samantha Piekos, Natalie Diaz, Jennifer Chen, Julia Wu, Prem Ramaswami, and James Manyika. 2023. [Data commons](#). *arXiv preprint arXiv:2309.13054*.
- Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FACt)*, pages 375–385.
- Krzysztof Janowicz, Armin Haller, Simon J. D. Cox, Danh Le Phuoc, and Maxime Lefrançois. 2019. [SOSA: A lightweight ontology for sensors, observations, samples, and actuators](#). *Journal of Web Semantics*, 56:1–10.
- Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. [Studying large language model behaviors under context-memory conflicts with real documents](#). In *First Conference on Language Modeling*.
- Hermann Kroll, Jan-Christoph Kalo, Denis Nagel, Stephan Mennicke, and Wolf-Tilo Balke. 2020. [Context-compatible information fusion for scientific knowledge graphs](#). In *Digital Libraries for Open Knowledge*, pages 33–47. Springer International Publishing.
- Maurizio Lenzerini. 2002. [Data integration: A theoretical perspective](#). In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246. Association for Computing Machinery.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5783–5797. Association for Computational Linguistics.
- Our World in Data. 2025. [Territorial and consumption-based CO2 emissions](#). Our World in Data Grapher. Data based on the Global Carbon Budget; accessed 2026-07-23.
- Rui Qiu, Shijie Chen, Yu Su, Po-Yin Yen, and Han Wei Shen. 2025. [Completing a systematic review in hours instead of months with interactive AI agents](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31559–31593, Vienna, Austria. Association for Computational Linguistics.
- Prashanth Radhakrishnan, Jennifer Chen, Bo Xu, Prem Ramaswami, Hannah Pho, Adriana Olmos, James Manyika, and R. V. Guha. 2024. [Knowing when to ask: Bridging large language models and data](#). *arXiv preprint arXiv:2409.13741*.
- Erhard Rahm and Philip A. Bernstein. 2001. [A survey of approaches to automatic schema matching](#). *The VLDB Journal*, 10(4):334–350.
- Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang. 2024. [The use of generative search engines for knowledge work and complex tasks](#). *arXiv preprint arXiv:2404.04268*.
- Zhiyin Tan and Jennifer D’Souza. 2026. [Diagnosing structural failures in LLM-based evidence extraction for meta-analysis](#). *arXiv preprint arXiv:2602.10881*. Accepted at IRCDL 2026.

- United Nations Economic Commission for Europe. 2024. [Generic statistical information model \(GSIM\), version 2.0: User guide](#). Technical report, United Nations Economic Commission for Europe.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: A free collaborative knowledgebase](#). *Communications of the ACM*, 57(10):78–85.
- Anzhe Xie, Weihang Su, Yujia Zhou, Yiqun Liu, and Qingyao Ai. 2026. [Benchmarking LLM agents on meta-analysis articles from nature portfolio](#). *arXiv preprint arXiv:2606.17041*.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge conflicts for LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.
- Jiani Zhang, Sercan O. Arik, Cosmin Arad, Fatma Ozcan, and Alon Halevy. 2026. [An agentic approach to metadata reasoning](#). *arXiv preprint arXiv:2604.20144*.
- Xiaoqing Zheng, Stuart E. Madnick, and Xitong Li. 2012. [SPARQL query mediation over RDF data sources with disparate contexts](#). In *Proceedings of the WWW 2012 Workshop on Linked Data on the Web*, volume 937 of *CEUR Workshop Proceedings*. CEUR-WS.org.

A Pilot design, prompts, and annotation

Code, prompts, model responses, and annotations are available at <https://github.com/JanKalo/akbc2026-vision>.

Data. All values are total annual CO₂ emissions from fossil fuels and industry, excluding land-use change, in 2022, in MtCO₂ (Our World in Data, 2025). Territorial and consumption-based values, respectively, were: Australia 384.1 and 365.1; United Kingdom 311.1 and 502.8; France 295.3 and 437.4; Germany 667.8 and 838.8; Japan 1029.6 and 1201.5. The UK is higher than France under both frameworks, but its territorial value is lower than France’s consumption-based value. Between Australia and the UK the ranking reverses between frameworks. Japan is higher than Germany under both.

Passages. Each passage was a paragraph of around 200 words about one country under one framework, written as a data-release note or a policy brief. It contained the value sentence “According to the Global Carbon Budget figures published by Our World in Data, [country]’s carbon dioxide emissions in 2022 totalled [value] million tonnes.” The next sentence named or described the framework, or provided no operationalization information. The surrounding text described the country, the year, the source, and the public use of the figure. It contained no term from either framework’s definition, no statement about where emissions occur or to whom they are attributed, and no numerals other than the value and the year. All passages cite the same source. The sentences varied across conditions as follows:

- Territorial, named: “The figure is the country’s territorial total, also published as production-based emissions.”
- Territorial, described: “The figure counts the carbon dioxide released inside the country’s own borders, from fuel burned and industrial processes there, regardless of where the resulting goods and services are used.”
- Consumption-based, named: “The figure is the country’s consumption-based total.”
- Consumption-based, described: “The figure counts the carbon dioxide released, wherever in the world it occurred, to produce the goods and services used by the country’s residents, so it includes emissions from making imports and leaves out emissions from making exports.”
- No operationalization, both frameworks: “The figure is the annual total for the calendar year, taken from the 2025 data release and rounded to one decimal place.”

Framework information	Ranked without qualification	Ranked with qualification	Declined
<i>Gemma 4 31B IT</i>			
Named	6	0	0
Described	6	0	0
Omitted	6	0	0
<i>Qwen3.8 Max</i>			
Named	0	6	0
Described	0	5	1
Omitted	6	0	0
<i>Claude Opus 5</i>			
Named	0	3	3
Described	0	2	4
Omitted	6	0	0

Table 3: Mismatched responses by model and operationalization information. Each row of counts covers six responses: three country pairs and two framework assignments.

Requests. Mismatched requests contained one passage per country, with the territorial passage first. The two assignments swapped which country used each framework. Both writing styles were used with each framework equally often. Matched requests contained two passages under the same framework, with the framework described. Complete requests contained all four passages, with the framework named or described. This gives 54 mismatched, 18 matched, and 18 complete requests. The system instruction was “You are a research assistant. Answer the user’s question using the retrieved excerpts.” The user message was “Did [country 1] or [country 2] have higher total annual CO₂ emissions in 2022?” followed by “Retrieved excerpts:” and the numbered passages. Gemma 4 31B IT, Qwen3.8 Max (0902), and Claude Opus 5 answered each request once at temperature 0, with a completion limit of 4,096 tokens. Qwen and Opus used medium reasoning effort. No response was empty or truncated.

Annotation. Each mismatched response was labeled ranked without qualification, ranked with qualification, or declined (Section 3.1); error asserted (a value must be wrong); or override (own figures used). We judged responses as a whole: numerical rankings with comparability caveats counted as qualified unless the response withheld a country ranking. Requiring figures on the same accounting basis did not itself count as declining. Matched responses were labeled direct answer, needless qualification, or refusal; complete responses, both rankings presented, a single ranking

with or without qualification, error asserted, or a ranking combining values across frameworks. Two separate sessions of a Claude assistant (recorded as Fable 5.1) initially agreed on 86/90 labels; the author resolved four differences, then reviewed 13 selected responses and changed three saved labels. Table 3 reports mismatched responses by model.