LCA Data Quality and the Advent of Artificial Intelligence

LCA Discussion Forum 92: AI in LCA: Innovations, Applications, and Challenges

26 February 2026

Achille Laurent, Ph.D., Strategic Project Manager

Expectation

Dataset creation is extremely time- and resource-intensive.

AI can accelerate the process and make it scalable.

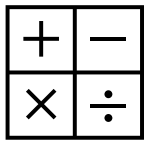
But at what cost?

"AI is not magic. It's a lot of math, a lot of data, and a lot of hard work."
— Jensen Huang, President and CEO, NVIDIA

Speaker Background

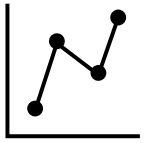
- PhD in Industrial Engineering: Operations Research — optimization of industrial processes and logistics networks
- 15+ years of experience in LCA and environmental quantification

Why this matters for today's talk:



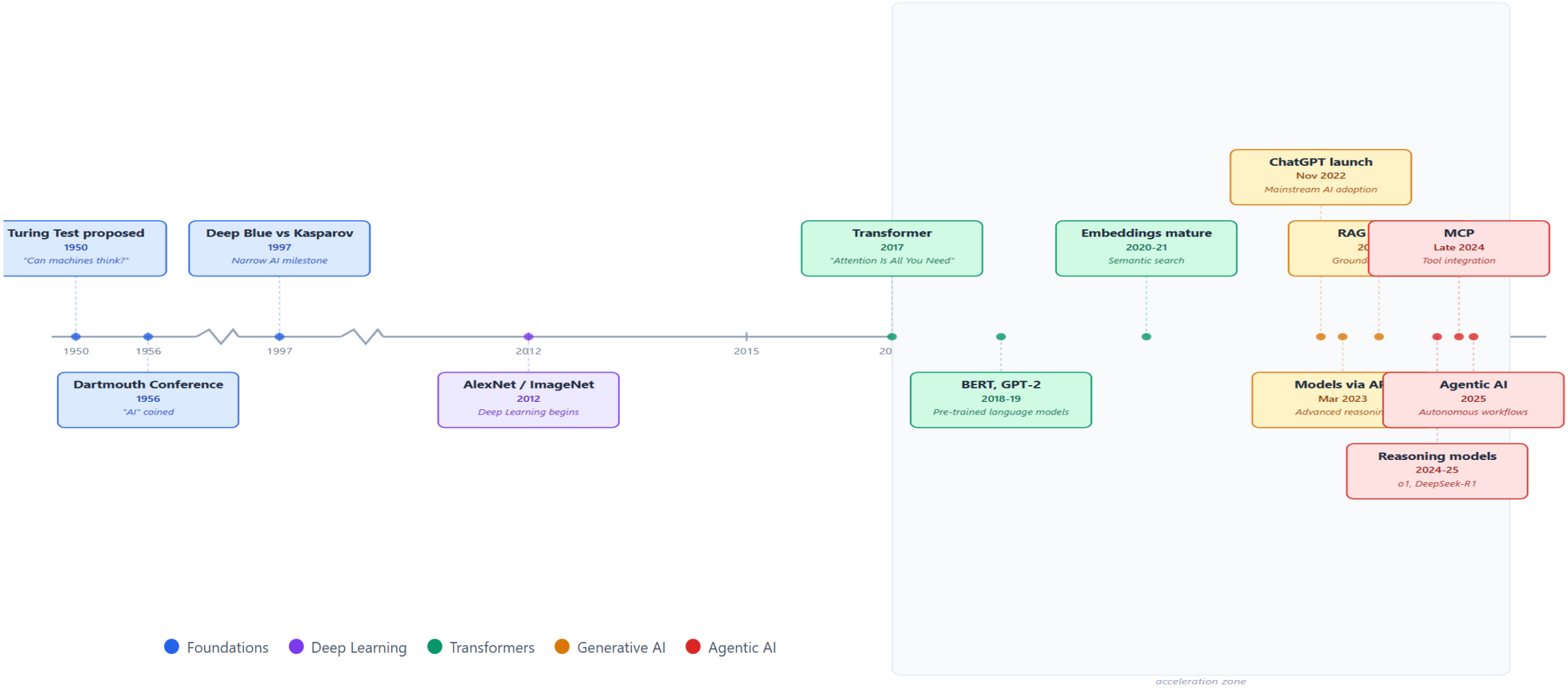
Optimization, not magic
It’s applied mathematics dealing with *vector distances, loss functions, constraints...*

Data quality, not hope
What LCA actually requires to be reliable



Modeling, not illusion
A model without statistics is incomplete and cannot be compared to anything

AI Evolution Timeline



Agentic AI (2025)

Agentic AI characteristics:

- Autonomous decision-making
- Tool use (search, code, databases)
- Multi-step reasoning
- Memory and context persistence

Risk amplification: Errors compound across autonomous steps

Data Quality

The ISO 14044 Mandate

Section 5.2.4

"Data quality requirements: specify in general terms the characteristics of the data needed for the study.

Descriptions *of data quality are important to understand the reliability of the study results and* ***properly interpret the outcome of the study****.“*

Section 4.1.6

*"**Transparency** — Due to the inherent complexity in LCA, transparency is an important guiding principle in executing LCAs, in order to **ensure a proper interpretation of the results**."*

ISO 14040:2006

→ **Data quality documentation and transparency serve the same purpose in ISO 14040: enabling proper interpretation of LCA results.**

The ISO 14044 Mandate

ISO 14044 Definition:

"Open, comprehensive, and understandable presentation of information"

Required documentation:

- Technology description
- Representativeness (geography, time, technology)
- Cut-off rules
- Uncertainty assessment

Operationalization: Pedigree Matrix

- Reliability
- Completeness
- Temporal, geographical, and technological correlation

Transparency as a Cornerstone

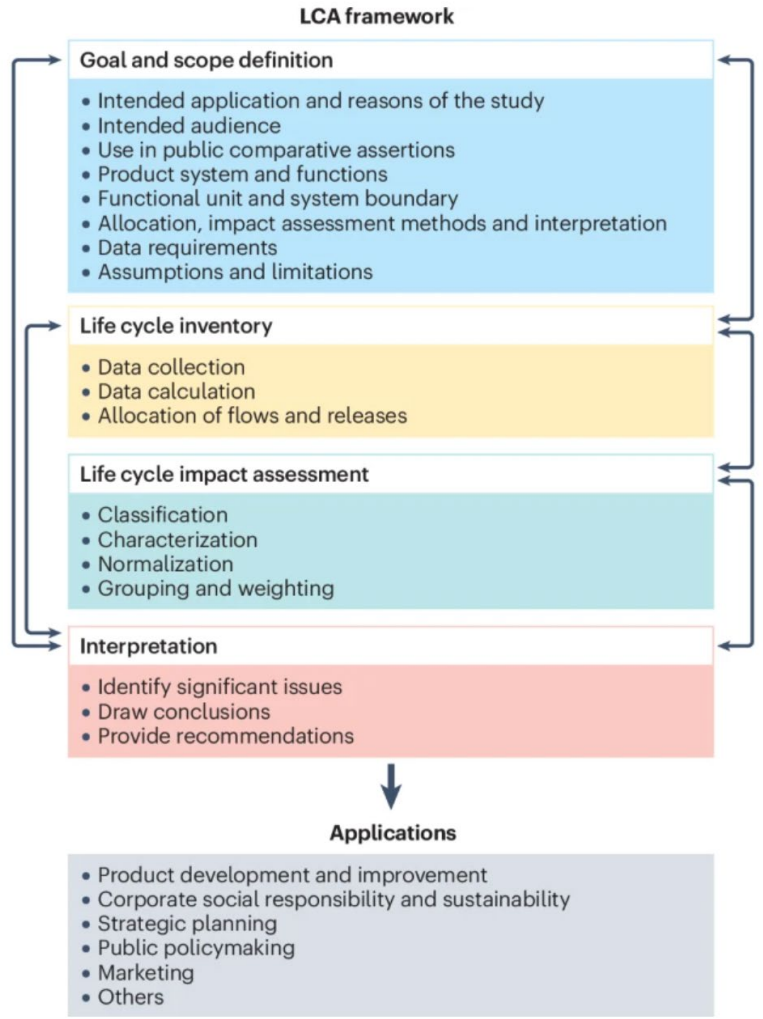
Why it matters:

- LCA is fundamentally a scientific endeavor
- Phase 4 of ISO 14040-44 = **Interpretation**
- Without proper documentation,
  sources/references, and
  uncertainty (pedigree matrix)

→ sensitivity and uncertainty analysis are impossible

- Large organizations are now audited on LCA reports

→ Traceability is mandatory



Xu et al. 2026

The FAIR Principles

Principle	Requirement
Findable	Sources have complete citation info (organization, author, title, year)
Accessible	Sources can be accessed via their reference information
Interoperable	Data can be integrated with other datasets and systems
Reusable	Data is well-documented for future use and adaptation

Inspired by Wilkinson *et al.* (2016)

The Transparency Challenge:

"The lack of transparency undermines LCA credibility and necessitates reconstructing databases for enhanced traceability... These findings highlight the challenge of data transparency and provide a methodology to evaluate databases."

- Guo et al. (2025)

The Transparency Challenge:

ecoinvent Performance in Comparative Study (Guo et al. 2025):

Metric	ecoinvent Result
Cross-citation complexity	Lowest among databases studied
Process transparency	Highest - most processes with findable & accessible sources

Why This Matters:

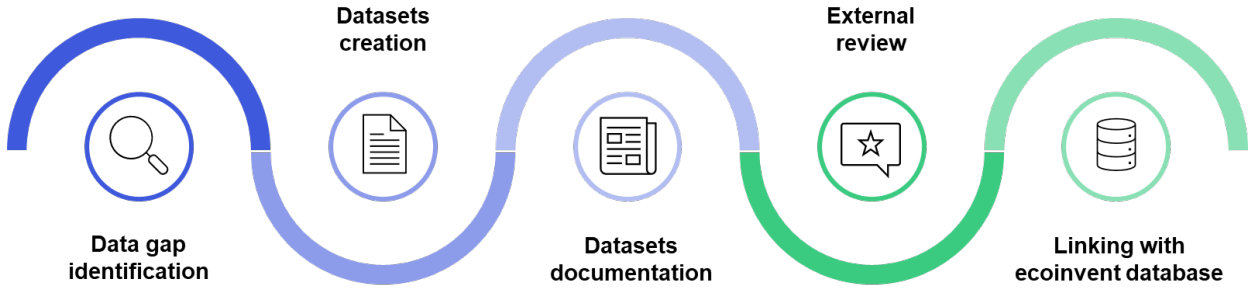
- Lowest complexity = easier traceability
- Highest transparency = sources are findable AND fully accessible

Transparency at ecoinvent

The ecoinvent Standard: *"In Transparency We Trust"*

This commitment means:

- No black-box calculations
- All assumptions documented
- Sources verifiable
- Uncertainty quantified and explained



ecoinvent's transparency comes from a rigorous, expert-driven process.

Can AI automation deliver the same guarantees?

The AI-LCA Market

Dominant Marketing Themes:

⚡ Speed: "Weeks not months", "real-time"

🤖 Automation: "Automate LCAs", "automate gap filling"

🎯 Simplicity: "Built for non-experts", "intuitive"

✅ Trust: "Audit-ready", "numbers you can trust"

💰 Cost: "Reduce costs", "€50 per EPD“

⚠️ **What's Rarely Discussed:**

- Source traceability and documentation
- Pedigree matrix / uncertainty quantification
- AI hallucination risks
- Need for expert validation
- ISO 14044 compliance of automated outputs

The AI-LCA Market

Dominant Marketing Themes:

⚡ Speed: "Weeks not months", "real-time"

🤖 Automation: "Automate LCAs", "automate gap filling"

🎯 Simplicity: "Built for non-experts", "intuitive"

✅ Trust: "Audit-ready", "numbers you can trust"

💰 Cost: "Reduce costs", "€50 per EPD“

⚠️ **What's Rarely Discussed:**

- Source traceability and documentation
- Pedigree matrix / uncertainty quantification
- AI hallucination risks
- Need for expert validation
- ISO 14044 compliance of automated outputs

Fast, Cheap, Good
Pick Two!

The Inevitable Challenge: Hallucinations

Definition:

Outputs that are plausible-sounding but factually incorrect, unsubstantiated, or nonsensical

"Hallucination is an innate and inevitable limitation of Large Language Models“ (Xu *et al.*, 2024)

Not a bug - a feature of the architecture:

- LLMs generate statistically probable text sequences
- They do NOT verify factual accuracy
- Cannot learn all computable functions
- Will inevitably produce outputs inconsistent with ground truth

Why Hallucinations Break LCA

The Problem Chain:

1. LLM prompted for inventory → Generates plausible but fabricated value
2. Output presented with high confidence
3. No traceable source

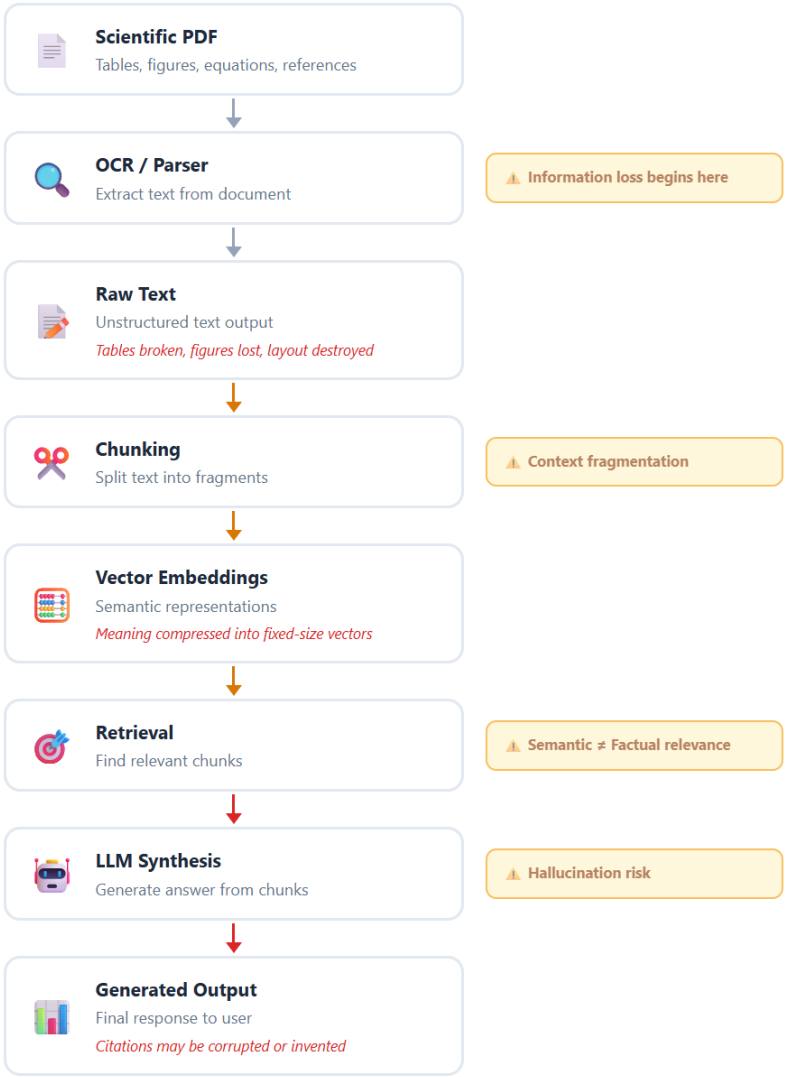
→ Fails pedigree matrix criteria:

❌ Reliability: Cannot verify
❌ Completeness: May be partial/invented
❌ Temporal correlation: Unknown
❌ Geographical correlation: Unknown
❌ Technological correlation: Unknown

So where exactly do these hallucinations come from?

Let's trace the pipeline

The Document Pipeline Challenge



AI Limitations - Document Extraction

1. Tables and Structured Data

"Current VDU methods suffer from OCR error propagation to the subsequent process.“ (Kim *et al.* 2022)

"The PDF format leads to a loss of semantic information, particularly for mathematical expressions.“ (Blecher *et al.* 2023)

AI Limitations - Document Extraction

1. Tables and Structured Data

2. Charts and Diagram Structure Loss

"State-of-the-art vision-language models do not perform well on [charts/plots].“ (Liu *et al.* 2022)

"Most methods... do not attempt to explicitly model the structure of the charts (e.g., how data is visually encoded and how chart elements are related to each other).“ (Masry *et al.* 2023)

AI Limitations - Document Extraction

1. Tables and Structured Data

2. Charts and Diagram Structure Loss

3. Vision-Text Inconsistency on Complex Tasks

"The trustworthiness of results derived from the vision modality diminishes as the tasks become more challenging.“ (Zhang *et al.* 2023)

AI Limitations - Document Extraction

1. Tables and Structured Data

2. Charts and Diagram Structure Loss

3. Vision-Text Inconsistency on Complex Tasks

For LCA:

- Inventory data in complex multi-column tables — extraction errors
- System boundaries in flow diagrams — visual information lost
- Actual data in Excel supplementary files — not parsed
- Units corrupted (MJ/FU)
- Multi-page tables — headers lost, rows mixed
- Footnotes with assumptions/sources — dissociated from content

AI Limitations - Citation Hallucinations in RAG

Step 1: Vector DB retrieval ✓ (correct chunks retrieved)

Step 2: LLM synthesis ✗ (citations corrupted/invented)

Why This Happens:

"Arbitrary and incorrect generations, termed 'confabulations,' are sensitive to irrelevant details and cannot be easily traced back to a verifiable source“ (Farquhar *et al.* 2024)

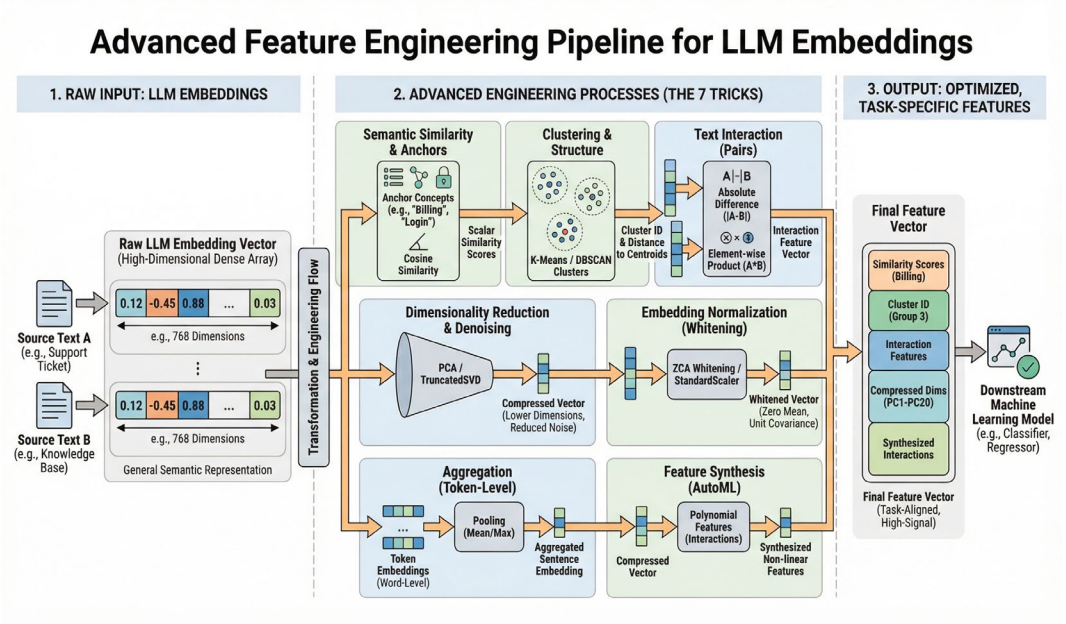
Types of Citation Corruption:

- Source mixing: Attributes content from Source A to Source B
- Fabricated details: Invents page numbers, years, author names
- Confident confabulation: Presents invented citations with certainty

Mitigation Strategies

For Document Extraction Issues:

1. **Use specialized parsers** (Nougat, Donut) before LLM processing
2. **Preserve table structure** with dedicated table extraction tools
3. **Process figures separately** with vision models, then integrate
4. **Maintain metadata** (page numbers, section headers) throughout pipeline
5. **Advanced embedding strategies**



Source: machinelearningmastery.com

Mitigation Strategies

For Document Extraction Issues:

For Citation Hallucinations:

1. **Require direct quotes** with full citation
2. **Cross-validate** with multiple sources
3. **Implement citation verification** as separate pipeline step with DOI resolution
4. **Use retrieval-only modes** when possible (no synthesis)

Mitigation Strategies

For Document Extraction Issues:

For Citation Hallucinations:

For LCA Specifically:

1. **Define system boundaries explicitly** before extraction
2. **Flag uncertainty** at each data point
3. **Maintain audit trail** from source to final value
4. **Expert review** remains mandatory for publication

Opening — The Road Ahead

Complex solutions on the horizon:

– **Adversarial Cooperation (Multi-Agent Verification)**

Multiple AI agents working together with different roles:

– One **proposes** data/values
– One **critiques** and challenges
– One **verifies** against sources
– One **synthesizes** final output

Benefits: Built-in error checking, transparent reasoning chains, reduced single points of failure (Block, Inc. 2025)

Opening — The Road Ahead

Complex solutions on the horizon:

- **Adversarial Cooperation (Multi-Agent Verification)**
- **Blockchain for data traceability**
- **Standardized global LCA data infrastructure**
- **AI-human hybrid workflows**

(Xu *et al.* 2025)

Opening — The Road Ahead

Complex solutions on the horizon:

- **Adversarial Cooperation (Multi-Agent Verification)**
- **Blockchain for data traceability**
- **Standardized global LCA data infrastructure**
- **AI-human hybrid workflows**

(Xu *et al.* 2025)

But a new paradox emerges: The environmental cost of environmental assessment

- LLM training: massive energy and water consumption
- Inference at scale: growing data center footprint
- Hardware lifecycle: rare earth extraction, e-waste

AI-assisted, not AI-driven

For LCA and data quality — human expertise remains essential (so far...)

Key Takeaways:

1. AI is transformative for LCA data collection speed
2. Fundamental limitations exist in document parsing and citation accuracy
3. ISO 14044 transparency requirements are non-negotiable
4. Human oversight remains essential for data quality

References

Blecher et al. (2023). Nougat: Neural Optical Understanding for Academic Documents. arXiv:2308.13418. https://doi.org/10.48550/arXiv.2308.13418

Block, Inc. (2025). Adversarial Cooperation in Code Synthesis. https://block.xyz/documents/adversarial-cooperation-in-code-synthesis.pdf

Farquhar et al. (2024). Detecting hallucinations in large language models using semantic entropy. Nature. https://doi.org/10.1038/s41586-024-07421-0

Guo et al. (2025). Shedding light on the shadows: LCA database transparency analysis. Journal of Industrial Ecology. https://doi.org/10.1111/jiec.70010

ISO 14040:2006. Environmental management — Life cycle assessment — Principles and framework.ISO 14044:2006. Environmental management — Life cycle assessment — Requirements and guidelines.

Kim et al. (2022). OCR-free Document Understanding Transformer (Donut). ECCV 2022. arXiv:2111.15664. https://doi.org/10.48550/arXiv.2111.15664

Liu et al. (2023). MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering. ACL 2023. arXiv:2212.09662. https://doi.org/10.48550/arXiv.2212.09662

Masry et al. (2023). UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning. arXiv:2305.14761. https://doi.org/10.48550/arXiv.2305.14761

Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Xu et al. (2025). Addressing critical challenges towards a robust data system for life cycle assessment. Nature Reviews Clean Technology, 1, 788-800. https://doi.org/10.1038/s44359-025-00107-4

Xu, et al. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817.https://doi.org/10.48550/arXiv.2401.11817

Zhang et al. (2023). Lost in Translation: When GPT-4V Can't See Eye to Eye with Text. arXiv:2310.12520. https://doi.org/10.48550/arXiv.2310.12520

Thank you.
Let's discuss!

Achille Laurent, Ph.D.
laurent@ecoinvent.org

Contact ecoinvent support