

How much metal does AI inference bind? Verification, uncertainty and a 2026 update of the MetallKIDD estimation tool

Severin Hackspiel

hackspiel.studio

ABSTRACT

Estimates of AI's environmental footprint mostly end at energy and carbon. The MetallKIDD tool, published in 2024, goes one step further: it translates the use of an AI service into inference GPUs and servers, and these into the mass of fifteen metals from the EU list of critical raw materials. This paper asks how robust that estimate is and how its 2023 assumptions hold up against data published until 2026. We rebuilt the tool's calculation in a script that reproduces all of its core values and its full scenario matrix, and found errors that move the main result: correcting them raises cumulative server metal for the default case (BingGPT on smartphones in Germany, 2023–2050) by 15 %, from 23.6 to 27.2 t, while an undocumented component mass alone could halve it. A sensitivity analysis shows that the choice between the tool's four calculation methods spans a factor of 275; across 10,000 Monte Carlo runs over documented ranges, the median is 3.1 t and the 90 % interval 0.07–23.6 t. Since 2023, measured energy per plain query has fallen to about a tenth of the value the tool assumed, while long reasoning queries match it; recalculated for 2026, the result therefore depends mainly on the share of reasoning. High-bandwidth memory, which the tool does not count, has become the dominant item in accelerator manufacturing. Fourteen of the fifteen metals are critical in the EU assessment, and recycled material covers little of their input.

Keywords: critical raw materials; AI inference; data centres; high-bandwidth memory; sensitivity analysis; recycling; rebound

This working paper is not a publication of the Wuppertal Institute and has not been peer reviewed. It examines the MetallKIDD tool (Austrup et al. 2024), of which the author is a co-author, developed in a project led by the Wuppertal Institute. Values from the tool refer to cells of the published file; all other values are taken from the sources in Appendix A.

1 Introduction

MetallKIDD (“Design eines zirkulären Metallmanagements mittels Metallstrategie und KI-Unterstützung für ein digitales Deutschland”) was a research project led by the Wuppertal Institute with Fraunhofer IOSB, the Folkwang University of the Arts and the VDI Technology Center, funded by the German Federal Ministry for the Environment (Bickel et al. 2024). One of its outcomes is a spreadsheet tool that estimates the future metal demand caused by the use of AI, published in June 2024 as a draft, version 0, under a CC BY 4.0 licence (Austrup et al. 2024). The tool was built as an estimate on best guesses, with every input open to change; it does not describe one scenario but shows which settings lead to which results.

This paper asks three questions. How much metal does the inference hardware of a single AI service bind according to the tool, and does the tool calculate what it intends to? How sensitive is the result to its inputs? And how do the assumptions of 2023 compare with data published until 2026? Its contributions are a verified reimplementations of the tool with an itemised list of errors and their effect (sections 3.3 and 4.2), a sensitivity and Monte Carlo

analysis (4.3), an update of energy, hardware and use with sourced values, including a recalculation for 2026 and a functional unit per 1,000 queries (4.4), and a criticality and recycling view per metal (4.5).

2 Related work

Estimates of the energy footprint of machine learning began with training (Strubell et al. 2019; Patterson et al. 2021, 2022) and have since shifted to inference, which dominates once models are deployed at scale (Wu et al. 2022; Desislavov et al. 2023). Measurements show that energy per task varies by orders of magnitude between model types and tasks (Luccioni et al. 2023, 2024). Recent per-query figures differ mainly in their system boundary: a provider-side measurement reports 0.24 Wh for the median Gemini Apps text prompt, including idle capacity, host processors and memory and data-centre overhead (Elsworth et al. 2025), while a bottom-up model estimates a median of 0.31 Wh for frontier-scale models on H100 nodes (Oviedo et al. 2026). At system level, supply-side projections (de Vries 2023) and scenario analyses (IEA 2025, 2026) put global data-centre electricity at around 950 TWh in 2030, within wide ranges.

Embodied impacts have received less attention. Gupta et al. (2021) showed that manufacturing dominates the carbon footprint of many computing systems. Life-cycle assessments of AI accelerators now exist from manufacturers (NVIDIA 2025b; Schneider et al. 2025); Schneider et al. attribute more than a third of embodied emissions to memory. Multi-criteria assessments go beyond carbon: Falk et al. (2026) find that teardown data raise the mineral and metal depletion of A100-based training by a third, and Berthelot et al. (2024) apply abiotic depletion indicators to a generative AI service. At fleet level, Wang et al. (2024) and de Vries-Gao (2026) estimate e-waste from AI servers, the latter revising earlier figures downward by anchoring them in production capacity, and Amoah et al. (2026) model demand for 20 materials in AI data-centre infrastructure, dominated by copper. Public inventories of servers (Talens Peiró and Ardente 2015) predate accelerators with high-bandwidth memory.

Criticality is assessed through supply risk and economic importance or vulnerability (Graedel et al. 2012; Blengini et al. 2017; Grohol and Veeh 2023); dynamic approaches follow how criticality changes as a technology scales (Roelich et al. 2014), and Schrijvers et al. (2020) note that data uncertainty is rarely carried into results. Rebound research cautions that efficiency gains per query need not reduce total demand (Coroamă and Mattern 2019; Lange et al. 2020; Luccioni et al. 2025). Methodological reviews show that top-down extrapolations of data-centre energy have repeatedly overstated growth and rest on poorly documented sources (Masanet et al. 2020; Mytton and Ashtine 2022); a review of AI life-cycle studies finds the same reliance on coarse proxies and little attention to materials (Lambert and Luccioni 2026).

We found few studies that translate a unit of inference use into metal masses, and those that model materials rarely itemise the memory that now makes up a growing share of accelerator hardware. The MetallKIDD tool was built for this link. This paper checks how well it holds and where it would need to change.

3 Methods

3.1 System boundary and functional unit

The system is the inference hardware of one AI service in one country: GPUs and the servers that host them, and the allocated share of end devices that run part of the inference. Training, networking, storage systems, cooling plant, buildings and grid connections are outside the boundary; so is memory as a component (section 4.4). Metal means the summed mass of fifteen metals (antimony, beryllium, boron, gallium, indium, cobalt, magnesium, manganese, neodymium, palladium, silicon, strontium, tantalum, titanium, tungsten), not the full mass of the hardware. The default case is BingGPT on smartphones in Germany from

2023 to 2050. As functional unit we add metal and energy per 1,000 queries; metal is allocated over a service life of $1/\rho$ years, where ρ is the renewal rate.

3.2 Calculation chain

With N AI users and a queries per user and day, the query rate is $q = N a / 86,400$ s. The tool estimates the number of GPUs in four ways (in the corrected form):

$$G_1 = q C_I / P_{\text{GPU}}$$

$$G_2 = q \tau P_{\text{dev}} / P_{\text{GPU}}$$

$$G_3 = q \cdot 3,600 e / W_{\text{GPU}}$$

$$G_4 = P_{\text{tot}} / (P_{\text{idle}} \text{ PUE}) - q (P_{\text{busy}} - P_{\text{idle}}) / (T P_{\text{idle}})$$

with compute per query C_I , GPU throughput P_{GPU} , the time τ and throughput P_{dev} of a reference device, energy per query e in Wh, GPU power W_{GPU} , and an accelerator model with total, idle and busy power and T queries per second per accelerator. The mean is increased by spare capacity o and reduced by the share ϕ carried by end devices, $G = \bar{G} (1 + o)(1 - \phi)$, and divided by n GPUs per server. Users grow with a damped adoption rate. From the start year, new hardware follows user growth with hardware per user falling by the efficiency rate ϵ per year; each year a share ρ of the installed stock is replaced by fewer, newer units in a vintage model. Metal is the sum over new servers and GPUs of their per-unit content w_m . Table 1 lists the inputs; the full equations are documented with the code (Appendix B).

Table 1. Main inputs, default value in the tool and range used in section 4.3. [A]: range set by us without a source.

Input	Default	Range, basis
Smartphone users, DE 2023	63.2 M	tool projection
Initial AI user share s	20 %	20–67 %; Eurostat 2025, Bitkom 2025
Queries per user and day a	15	7.5–22.5 [A]
Energy per query e	3.71 Wh	0.24–6.8 Wh; Elsworth et al. 2025, tool
Compute per query C_I	3,956 TFLOP	4 GPUs $\times$ 1 s $\times$ 989 TFLOP/s
GPU power W_{GPU}	700 W	700–1,000 W; NVIDIA
Calculation method	mean	methods 1–4, mean, median
Spare capacity o	50 %	0–100 % [A]
Device load for inference	10 %	10–20 %; tool
GPUs per server n	8	–
Renewal rate ρ	20 %/a	0–30 %; tool scenarios
Efficiency rate ϵ	5 %/a	0–30 %; tool scenarios
Server periphery	30 %	0–30 %; not applied in v0
Main board mass	4,000 g	265–4,000 g; section 4.2
Metal per server / GPU	632.6 / 17.3 g	tool component data

3.3 Verification and correction

We reimplemented the chain in Python, including the four methods, the vintage model for servers and GPUs, the allocation to end devices, recycling and the criticality comparison, and compared it with the cached values of the published workbook. The formulas of all 148 sheets were then reviewed for broken references, typed-in values where formulas are expected, inconsistent formulas within a column and unit errors. Each error was corrected separately and in combination to quantify its effect.

3.4 Uncertainty analysis

On the corrected model, we varied each input of table 1 over its range with all others at their defaults (one at a time) and ran 10,000 Monte Carlo draws: energy per query log-uniform, the calculation method uniform over its six options, all other inputs uniform. Target values are cumulative server and GPU metal 2023–2050 and the number of GPUs in 2023. The distributions are our assumption; they express ignorance, not knowledge of the true spread.

3.5 Data for the update

Values for 2026 were taken only from sources that could be retrieved and checked on 3 October 2026. Appendix A classifies each source by type (peer-reviewed, agency, company statement, market research) and states whether it reports a measurement or a projection. Company statements are used where no independent measurement exists and are marked as such.

4 Results

4.1 The tool as published

For the default case, Germany has 12.6 million AI users in 2023 (20 % of 63.2 million smartphone users), producing 2,196 queries per second. The four methods give between 153 and 41,895 GPUs (figure 1); their mean, with 50 % spare capacity and after 11.2 % carried by phones, gives 17,270 GPUs in 2,159 servers containing 1.67 t of the tracked metals, 1.29 t of it silicon. With 20 % renewal, 46.5 t are bought until 2050 without efficiency gains and 23.6 t with 5 % per year (figure 2); across the tool's scenarios the range is 3.6 to 65.7 t (figure 3).

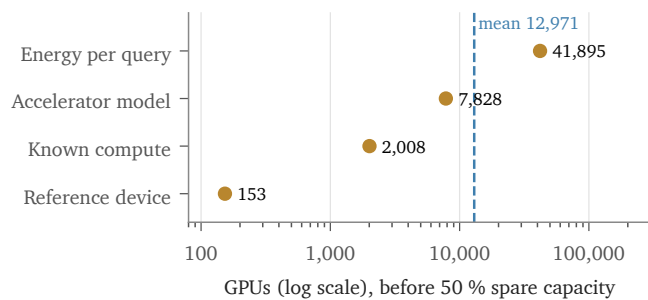


Figure 1. GPUs for the default case by calculation method in version 0, before spare capacity. Method 1 is corrected to 8,783 in section 4.2.

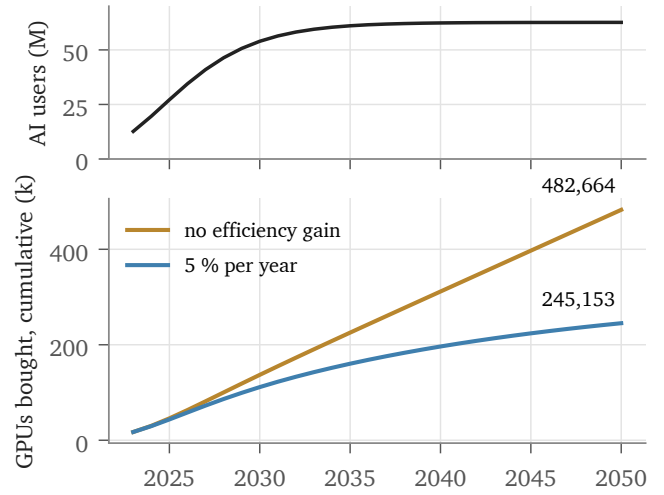


Figure 2. AI users (top) and GPUs bought since 2023 (bottom) with 20 % renewal, without and with 5 % efficiency per year; version 0.

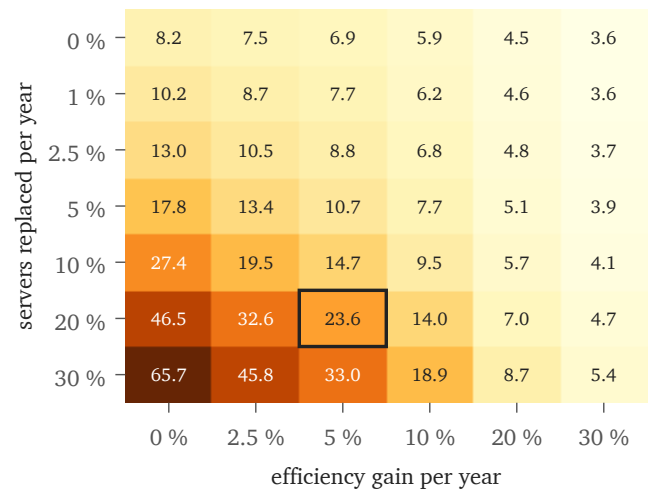


Figure 3. Server and GPU metal cumulated 2023–2050 in tonnes by renewal and efficiency rate, version 0. Framed: default. The 40 % column is omitted because version 0 computes it only partly with that rate.

4.2 Verification and errors

The reimplementing reproduces the core values of the workbook exactly, among them 1.6655 t in 2023, 46.547 t and 23.642 t cumulated, 17,270 GPUs and the 11.24 % device share, and all 49 cells of the renewal × efficiency matrix (largest deviation $4 \cdot 10^{-14}$ %). Small deviations of up to 0.5 % in methods 1 to 3 come from rounded numbers typed into the workbook.

The review found errors of three kinds. Some change the main result (table 2). Method 1 divides the time per query by the number of GPUs a model instance needs (350 GB model on 80 GB GPUs) instead of counting them, which gives 2,008 instead of 8,783 GPUs. The growth of device compute starts in 2021 instead of the start year, which inflates the device share. The 30 % server periphery is entered but not applied. The main board, which carries 74 % of server metal, is calculated with 4,000 g, while the documentation lists 264.97 g with the note “prüfen!” (to be checked); according to the author, that value was taken from the board of a consumer graphics card and does not

describe a server main board, and the 4,000 g have no documented source. We keep 4,000 g and treat the range in section 4.3.

Table 2. Effect of corrections on cumulative server and GPU metal 2023–2050 (20 % renewal, 5 % efficiency) and on GPUs in 2023. Version 0: 23.64 t, 17,270 GPUs.

Correction	Metal	GPUs
Method 1 counts GPUs per instance	+13.0 %	+13.1 %
Device compute grows from start year	+1.7 %	+1.7 %
Main correction, combined	27.18 t, +15.0 %	19,862
+ periphery 30 % applied	33.87 t, +43.2 %	19,862
+ main board 265 g instead of 4,000 g	11.85 t, -49.9 %	19,862
+ periphery and 265 g	13.94 t, -41.0 %	19,862

Other errors change secondary results. The device time series is offset by two years; corrected, cobalt allocated to phones in 2023 falls from 38.65 to 15.55 t (−60 %). Palladium is recycled at the rate for neodymium from 2024 on, which understates secondary palladium twenty-fold. In the criticality sheet, one column multiplies by the power of the scaled-up ecosystem in gigawatts instead of a device factor, and the comparison adds cumulative server purchases to a device stock. The energy of method 4 is calculated with a factor of 60 instead of 3,600. Method 4 falls as demand rises and turns negative above about 58 million users, and AI users are not capped at the number of smartphone users.

A third group explains why the published file does not react as intended: 6,801 cells show reference errors, partly because sheet names were cut off at 31 characters; several intermediate results are typed in, so that the selection of a calculation method, the adoption rate, energy per query and queries per day do not change the hardware result; and the 25 % and 40 % efficiency scenarios fall back to 20 % and 30 % for growth. Data issues include a GPU throughput of 989 TFLOP/s that NVIDIA lists with sparsity, a training time of 168 h where the cited source reports 14.8 days (Patterson et al. 2021), smartphone composition data without a source and with a physically implausible hydrogen content, and server silicon that stems mostly from the glass fibre of circuit boards rather than from semiconductors, while it is compared with EU demand for silicon metal.

4.3 Sensitivity

The choice of calculation method dominates (figure 4): from method 2 to method 3, cumulative metal ranges from 0.28 to 77.6 t, a factor of 275. Next follow the efficiency rate (5.4–53.5 t), energy per query (9.0–43.3 t), the renewal rate (7.9–37.9 t) and queries per day (14.2–39.8 t). The main board mass spans a factor of 2.3, spare capacity 2.0, GPU power, periphery, user share and device load less than 1.3. The user share matters for the hardware in 2023

but little for metal until 2050, because user numbers reach the number of smartphones in either case.

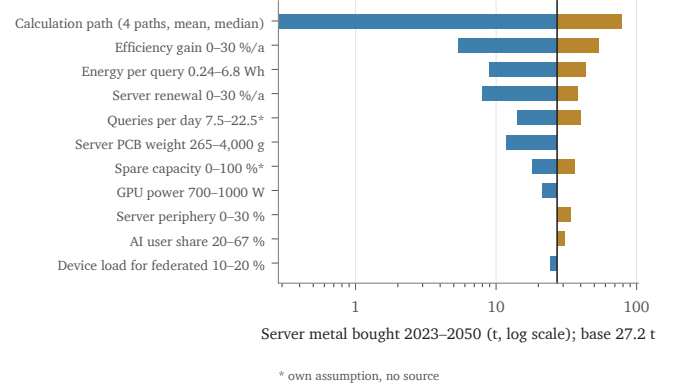


Figure 4. One-at-a-time sensitivity of cumulative server and GPU metal 2023–2050, corrected model (base 27.2 t). [A]: range without a source.

Across 10,000 Monte Carlo draws with a random calculation method, the median is 3.1 t and the 90 % interval 0.07–23.6 t (figure 5); for GPUs in 2023 the median is 11,042 (281–65,514). With the method fixed to the mean, the median is 4.25 t (1.27–20.2 t). The corrected default of 27.2 t lies at about the 96th percentile, because it combines three inputs at the upper end of their ranges: a low efficiency rate, a high energy per query and the higher main board mass. Rank correlations confirm efficiency ($\rho = -0.29$), energy per query (+0.22), renewal (+0.20) and main board mass (+0.19) as the main drivers after the method.

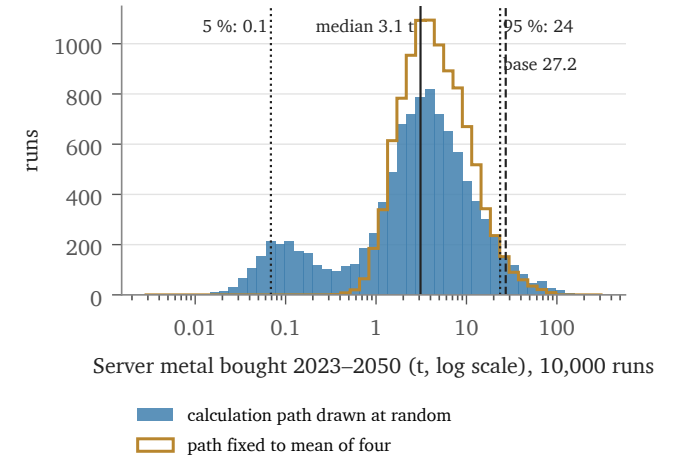


Figure 5. Distribution of cumulative server and GPU metal 2023–2050 over 10,000 draws. Blue: calculation method drawn at random, with 5th, 50th and 95th percentile; gold: method fixed to the mean.

4.4 Update 2026

Energy per query. Measured energy per plain query is now an order of magnitude below the tool's assumption (figure 6). Google reports 0.24 Wh for the median Gemini Apps text prompt in May 2025, a company measurement with a disclosed method (Elsworth et al. 2025). A peer-reviewed model finds a median of 0.31 Wh for frontier-scale models and 3.91 Wh for long reasoning queries, and concludes that widely cited estimates are overstated by 4 to 20 times

(Oviedo et al. 2026). The tool's 3.71 Wh thus describe a reasoning query rather than a plain one. Benchmarks of open models find reasoning variants using on average 30 times the GPU energy of models without reasoning (Luccioni and Gamazaychikov 2025). A query is no longer a stable unit.

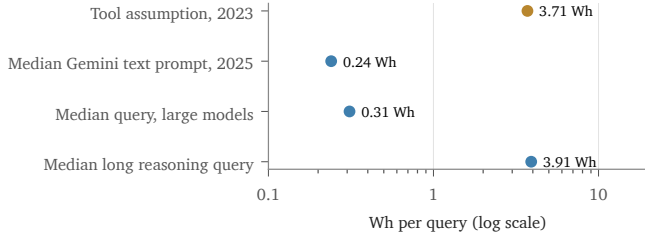


Figure 6. Energy per query: the tool's assumption against values published in 2025 and 2026.

Memory. The tool's server carries 8×80 GB of high-bandwidth memory; current eight-GPU systems carry 1,440 GB and 2.1 TB (NVIDIA 2025a). In NVIDIA's externally verified carbon footprint of an HGX B200 board, memory accounts for 49 % of 2,274 kg CO₂e from manufacturing (NVIDIA 2025b), consistent with Schneider et al. (2025). The IEA (2026) expects the shortage of such memory to last until at least the end of 2027, and contract prices for conventional DRAM nearly doubled in the first quarter of 2026 (TrendForce 2026). A metal balance without memory misses the component that now dominates both cost and manufacturing footprint.

Recalculation for 2026. We reran the tool's energy method (method 3) with sourced values only: 0.31 Wh for a plain and 3.91 Wh for a reasoning query (Oviedo et al. 2026) and 1,000 W per GPU (NVIDIA 2025b). All other inputs stay as in the tool, including 34.5 million AI users in 2026 from its own projection, which lies between the one third of EU residents aged 16 to 74 who used generative AI in 2025 (Eurostat 2025) and a survey value of 67 % (Bitkom 2025). The metal content per server remains that of the tool, since no material inventory of current servers is published; this mixes 2026 hardware with 2023 composition and is a limitation, not a result.

Table 3. Energy method (method 3), version 0 inputs and 2026 cases, with the functional unit per 1,000 queries. Metal per 1,000 queries is server and GPU metal allocated over five years of use.

	2023, v0	2026 plain	2026 reasoning
AI users	12.6 M	34.5 M	34.5 M
Wh per query / W per GPU	3.71 / 700	0.31 / 1,000	3.91 / 1,000
GPUs	55,779	8,891	112,146
Servers	6,972	1,111	14,018
Metal, first year	5.4 t	0.9 t	10.8 t
Electricity per year	257 GWh	59 GWh	738 GWh
Metal per 1,000 queries	15.5 mg	0.9 mg	11.5 mg
Energy per 1,000 queries	3,710 Wh	310 Wh	3,910 Wh

With plain queries, the service needs fewer GPUs than in 2023 despite almost three times the users; with reasoning queries only, twice as many. In between, the result depends linearly on the share of reasoning (figure 7), a parameter the tool does not have. Data centres in Germany used 21.3 TWh in 2025 (Hintemann and Hinterholzer 2025); the service would account for 0.3 to 3.5 % of that.

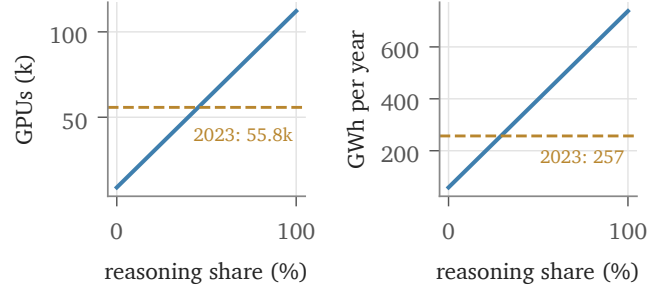


Figure 7. GPUs and electricity for Germany in 2026 by share of reasoning queries, energy method. Dashed: the same method with the 2023 inputs.

Electricity. Data centres worldwide used about 415 TWh in 2024 (IEA 2025) and 485 TWh in 2025; AI-focused data centres grew by 50 % in 2025, and the IEA (2026) projects 950 TWh by 2030, around 3 % of global electricity demand (figure 8). This growth alongside large efficiency gains per task is the pattern rebound research describes (Luccioni et al. 2025).

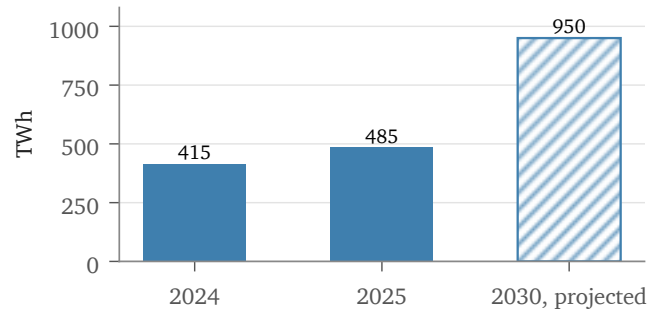


Figure 8. Electricity consumption of data centres worldwide; 2030 projected.

Devices and lifetimes. Phones now run their own models of about three billion parameters (Apple 2025), so the relevant split is between small local models and large ones in the cloud, not the share of one large model that a network of phones could carry. The tool's renewal rate holds: Microsoft and Alphabet depreciate servers over six years, Meta over five and a half, and Amazon shortened part of its servers from six to five years in 2025, citing the pace of AI (annual reports on Form 10-K). Five to six years correspond to 17–20 % renewal per year; depreciation is an accounting figure, not a measured service life.

4.5 Criticality and recycling

All fifteen metals except indium are critical in the EU assessment of 2023 (table 4). For nine of them a single country supplies more than half of world production; China produces 99 % of primary gallium (USGS 2026), and data

centres could need up to a tenth of today's gallium supply by 2030 (IEA 2025). The EU puts the recycled share of input at zero for gallium, beryllium, strontium and silicon metal; only tungsten reaches 42 %. Several rates in the

tool are well above the EU values, most clearly for titanium (91 against 1 %); for gallium, strontium and tantalum the tool combines an end-of-life recycling rate below 1 % with a recycled input of 25 %, which is inconsistent.

Table 4. The fifteen metals in the EU criticality assessment 2023 (Grohol and Veeh 2023): supply risk (SR), economic importance (EI), end-of-life recycling input rate (EOL-RIR), main supplier by share of world production; rate used in the tool; server and GPU metal cumulated 2023–2050, version 0; share of annual EU demand, corrected, with and without scaling to the tool's AI ecosystem. Thresholds: $SR \geq 1$, $EI \geq 2.8$. Neodymium and palladium recycling rates refer to light rare earths and platinum group metals.

Metal	SR	EI	EOL-RIR	Tool	Main supplier	Servers to 2050	EU demand	scaled
Neodymium	4.5	7.2	1 %	1 %	China 85 %	0.235 t	0.73 %	37 %
Magnesium	4.1	7.4	13 %	33 %	China 91 %	1.15 t	0.06 %	2.9 %
Gallium	3.9	3.7	0 %	25 %	China 94 %	0.061 t	0.05 %	2.7 %
Boron	3.6	3.9	1 %	0.6 %	Türkiye 48 %	3.13 t	<0.01 %	0.07 %
Cobalt	2.8	6.8	22 %	34 %	DR Congo 63 %	0.030 t	1.09 %	56 %
Strontium	2.6	6.5	0 %	25 %	Iran 37 %	0.121 t	<0.01 %	<0.01 %
Antimony	1.8	5.4	28 %	27 %	China 56 %	1.02 t	<0.01 %	0.29 %
Beryllium	1.8	5.4	0 %	7 %	USA 67 %	0.030 t	0.11 %	5.8 %
Titanium metal	1.6	6.3	1 %	91 %	China 43 %	0.013 t	<0.01 %	<0.01 %
Palladium	1.5	8.1	12 %	50 %*	Russia 40 %	0.061 t	0.07 %	3.6 %
Silicon metal	1.4	4.9	0 %	–	China 76 %	36.0 t	<0.01 %	0.10 %
Tantalum	1.3	4.8	1 %	25 %	DR Congo 35 %	2.44 t	0.13 %	6.5 %
Manganese	1.2	6.9	9 %	9 %	South Africa 29 %	2.27 t	<0.01 %	0.02 %
Tungsten	1.2	8.7	42 %	42 %	China 86 %	0.002 t	0.04 %	2.1 %
Indium	0.6	2.6	1 %	0.001 %	China 50 %	≈ 0	<0.01 %	0.03 %

* Entered as 50 %, calculated with 1 % from 2024 on. EU demand: annual average 2023–2050 of corrected server purchases plus allocated device purchases, divided by EU demand in the tool. Scaled: servers and devices multiplied by the tool's ecosystem factor of 51.15, which extrapolates from one service to all AI applications; this factor is itself an estimate of version 0.

For a single service the shares stay at or below about 1 % of annual EU demand, led by cobalt and neodymium in phones. Scaled to the tool's estimate of the whole AI ecosystem, cobalt reaches 56 % and neodymium 37 %. These figures should be read as an indication of where the pressure would fall, not as a forecast: they depend on the ecosystem factor, on smartphone data without a documented source and on the allocation of phones to AI use. How much e-waste AI hardware will leave is likewise disputed, between 1.2 and 5.0 million tonnes cumulated 2020–2030 from generative AI (Wang et al. 2024) and 131–225 kt per year from AI servers in 2030 (de Vries-Gao 2026).

5 Discussion

The tool's structure, from the compute behind a use to the metal in servers, holds, and its central idea, to make visible how far apart reasonable estimates lie, is confirmed by the analysis: the method choice alone spans more than two orders of magnitude. What the analysis changes is how the default result should be read. The published value of 23.6 t is not a central estimate; after correction it rises to 27.2 t but sits near the upper end of the distribution. A single undocumented component mass moves it as much as half.

The update shifts which parameters matter. Energy per plain query has fallen by an order of magnitude, while reasoning has opened a new spread of the same size; the share of reasoning would now rank among the most important inputs. Memory, absent from the tool, has become the largest item in the manufacturing footprint of accelerators. And the tool's efficiency rate, which it defines as hardware actually saved after rebound, meets a period in which efficiency per task rose sharply while data-centre electricity doubled; a version that wants to say something about rebound would have to model demand, not only efficiency.

6 Limitations

The recalculation for 2026 combines current energy and power values with the 2023 composition of servers and GPUs and without memory, because no material inventory of current AI servers is published. The 2026 user number is the tool's projection, and no source exists for queries per user and day. The ranges marked [A] and the Monte Carlo distributions are our assumptions. Several 2026 values are company statements (Appendix A). The device data of the tool lack a source. The paper was written by a co-author of the tool and has not been reviewed by others.

7 Conclusion and version 1

An estimate of AI's metal demand in 2026 would have to start from three points: the spread between calculation methods, the share of reasoning in use and the memory in the hardware. A version 1 of the tool would count tokens instead of queries and treat reasoning as its own parameter; add memory both as a constraint and as a component with its own materials; use a current inventory of servers including weight, cooling and power supply; calculate energy as a time series; model devices as running their own small models; align recycling rates with the EU assessment; replace typed-in values with formulas; and report results as ranges with the sensitivity shown here, rather than as a single default.

Acknowledgements and disclosure

The tool was developed with Vera Henrike Austrup and Manuel Bickel in the MetallKIDD project. The reimplementation, the formula review, the literature search and a first draft of sections 2 to 7 were carried out with the help of an AI system (Claude Opus 5.5, Anthropic); all values were checked against the published tool and the cited sources, and responsibility lies with the author. The author is a co-author of the tool examined here. No funding was received for this paper.

Data and code availability

The tool is available under CC BY 4.0 at doi:10.5281/zenodo.11506192. The reimplementation, the verification against the workbook and the sensitivity analysis are available as Python scripts at github.com/strhwste/metallkidd-analysis (Appendix B).

References

- Amoah, M., Brown, M., Simon, A., Bazilian, M., Matisek, J. (2026): Mineral demand from AI data centers. *Resources Policy* 119, 105970. doi:10.1016/j.resourpol.2026.105970
- Apple (2025): Updates to Apple's on-device and server foundation language models. Apple Machine Learning Research, 17 July 2025.
- Austrup, V. H., Hackspiel, S., Bickel, M. (2024): MetallKIDD-Tool, version 0. Zenodo. doi:10.5281/zenodo.11506192
- Berthelot, A., Caron, E., Jay, M., Lefèvre, L. (2024): Estimating the environmental impact of generative-AI services using an LCA-based methodology. *Procedia CIRP* 122, 707–712. doi:10.1016/j.procir.2024.01.098
- Bickel, M., Austrup, V., Hackspiel, S. et al. (2024): Kurzbericht zum Projekt MetallKIDD. Final report part I, FKZ 67KI2046. Wuppertal Institut. edocs.tib.eu/files/e01fb24/1910593397.pdf
- Bitkom (2025): KI-Nutzung boomt. Press release, 5 May 2025 (survey, n = 1,005).
- Blengini, G. A. et al. (2017): EU methodology for critical raw materials assessment. *Resources Policy* 53, 12–19. doi:10.1016/j.resourpol.2017.05.008
- Coroamă, V. C., Mattern, F. (2019): Digital rebound – why digitalization will not redeem us our environmental sins. *Proc. ICT4S 2019, CEUR-WS* 2382.
- de Vries, A. (2023): The growing energy footprint of artificial intelligence. *Joule* 7(10), 2191–2194. doi:10.1016/j.joule.2023.09.004
- de Vries-Gao, A. (2026): Recalibrating global artificial intelligence e-waste estimates. *Resources, Conservation and Recycling* 229, 108872. doi:10.1016/j.resconrec.2026.108872
- Desislavov, R., Martínez-Plumed, F., Hernández-Orallo, J. (2023): Trends in AI inference energy consumption. *Sustainable Computing* 38, 100857. doi:10.1016/j.suscom.2023.100857
- Elsworth, C. et al. (2025): Measuring the environmental impact of delivering AI at Google scale. arXiv:2508.15734. Summary: Vahdat, A., Dean, J., Google Cloud blog, 21 August 2025.
- Eurostat (2025): 32.7 % of EU people used generative AI tools in 2025. News item, 16 December 2025.
- Falk, S. et al. (2026): More than carbon: cradle-to-grave environmental impacts of GenAI training on the Nvidia A100 GPU. *Environmental Impact Assessment Review* 121, 108525. doi:10.1016/j.eiar.2026.108525
- Graedel, T. E. et al. (2012): Methodology of metal criticality determination. *Environmental Science & Technology* 46(2), 1063–1070. doi:10.1021/es203534z
- Grohol, M., Veeh, C. (2023): Study on the critical raw materials for the EU 2023. European Commission. doi:10.2873/725585
- Gupta, U. et al. (2021): Chasing carbon: the elusive environmental footprint of computing. *IEEE HPCA 2021*, 854–867. doi:10.1109/HPCA51647.2021.00076
- Hintemann, R., Hinterholzer, S. (2025): Rechenzentren in Deutschland: aktuelle Marktentwicklungen, Update 2025. Bitkom, Borderstep Institut.
- IEA (2025): Energy and AI. International Energy Agency, April 2025.
- IEA (2026): Key questions on energy and AI. International Energy Agency, April 2026.
- Lambert, K., Luccioni, S. (2026): From cradle to cloud: a life cycle review of AI's environmental footprint. *Proc. ACM FAccT 2026*, 518–540. doi:10.1145/3805689.3812401
- Lange, S., Pohl, J., Santarius, T. (2020): Digitalization and energy consumption. *Ecological Economics* 176, 106760. doi:10.1016/j.ecolecon.2020.106760
- Luccioni, A. S., Viguier, S., Ligozat, A.-L. (2023): Estimating the carbon footprint of BLOOM. *JMLR* 24(253), 1–15.
- Luccioni, S., Jernite, Y., Strubell, E. (2024): Power hungry processing. *Proc. ACM FAccT 2024*, 85–99. doi:10.1145/3630106.3658542
- Luccioni, A. S., Strubell, E., Crawford, K. (2025): From efficiency gains to rebound effects. *Proc. ACM FAccT 2025*, 76–88. doi:10.1145/3715275.3732007
- Luccioni, S., Gamazaychikov, B. (2025): AI Energy Score v2: refreshed leaderboard, now with reasoning. Hugging Face, 4 December 2025.
- Masanet, E., Shehabi, A., Lei, N., Smith, S., Koomey, J. (2020): Recalibrating global data center energy-use estimates. *Science* 367(6481), 984–986. doi:10.1126/science.aba3758
- Mytton, D., Ashtine, M. (2022): Sources of data center energy estimates. *Joule* 6(9), 2032–2056. doi:10.1016/j.joule.2022.07.011
- NVIDIA (2025a): DGX B200 and DGX B300, product specifications. Retrieved 3 October 2026.
- NVIDIA (2025b): Product carbon footprint summary for NVIDIA HGX B200. July 2025.
- Oviedo, F. et al. (2026): Energy use of AI inference, efficiency pathways, and test-time scaling. *Joule* 10(8), 102430. doi:10.1016/j.joule.2026.102430
- Patterson, D. et al. (2021): Carbon emissions and large neural network training. arXiv:2104.10350.
- Patterson, D. et al. (2022): The carbon footprint of machine learning training will plateau, then shrink. *Computer* 55(7), 18–28.
- Roelich, K. et al. (2014): Assessing the dynamic material criticality of infrastructure transitions. *Applied Energy* 123, 378–386. doi:10.1016/j.apenergy.2014.01.052
- Schneider, I. et al. (2025): Life-cycle emissions of AI hardware: a cradle-to-grave approach and generational trends. arXiv:2502.01671.
- Schrijvers, D. et al. (2020): A review of methods and data to determine raw material criticality. *Resources, Conservation and Recycling* 155, 104617. doi:10.1016/j.resconrec.2019.104617

- Strubell, E., Ganesh, A., McCallum, A. (2019): Energy and policy considerations for deep learning in NLP. Proc. ACL 2019, 3645–3650. doi:10.18653/v1/P19-1355
- Talens Peiró, L., Ardente, F. (2015): Analysis of material efficiency requirements of enterprise servers. JRC, EUR 27467 EN. doi:10.2788/409022
- TrendForce (2026): DRAM industry revenue, first quarter 2026. Press release, 1 June 2026.
- USGS (2026): Mineral commodity summaries 2026: gallium. doi:10.3133/mcs2026
- Wang, P., Zhang, L.-Y., Tzachor, A., Chen, W.-Q. (2024): E-waste challenges of generative artificial intelligence. Nature Computational Science 4(11), 818–823. doi:10.1038/s43588-024-00712-6
- Wu, C.-J. et al. (2022): Sustainable AI: environmental implications, challenges and opportunities. Proc. MLSys 4.

Appendix A: Sources of the 2026 values

Type, kind of value and reliability of the sources used for the update. All retrieved 3 October 2026.

Source	Type	Value	Reliability
Oviedo et al. 2026	peer-reviewed	model with measured inputs	high; authors at Microsoft
Elsworth et al. 2025	company, preprint	measurement, method disclosed	medium; not reproducible
Luccioni, Gamazaychikov 2025	benchmark, blog	measurement, GPU only	medium; open models only
NVIDIA 2025a	company datasheet	specification	medium; maximum values
NVIDIA 2025b	company, ISO 14067, verified	modelled LCA	medium–high for CO ₂ e
IEA 2025, 2026	intergovernmental agency	statistics, projection	high for system values
TrendForce 2026	market research	survey, proprietary	low–medium
Hintemann, Hinterholzer 2025	industry association study	model	medium
USGS 2026	agency	statistics	high
Eurostat 2025	official statistics	survey	high for user share
Grohol, Veeh 2023	European Commission	assessment, data 2016–2020	high
Wang et al. 2024	peer-reviewed	projection, cumulative	medium; revised by de Vries-Gao
de Vries-Gao 2026	peer-reviewed	projection, annual	high
Apple 2025	company	model description	medium
Form 10-K reports	statutory filings	accounting lives	high as accounting figure

Appendix B: Code

metallkidd_modell.py implements the chain of section 3.2 with all corrections as switches; verifikation.py compares it with the cached values of the workbook; analyse.py runs the sensitivity and Monte Carlo analysis (seed 20261003); wirkung.py quantifies each correction. The scripts read the published workbook and need Python 3 with openpyxl, numpy and matplotlib. Repository: github.com/strhwste/metallkidd-analysis.