REFERENCE DOCUMENT

Estimating the Inference Carbon Intensity of DeepSeek Models: Methodology, Estimates, and Justification

Prepared by: **InferenceCarbon.ai Reference Team**

Date: 19 March 2026 | Version: 1.4 | Classification: For publication — subject to stated caveats

Important note: *DeepSeek has published no per-query energy data, no sustainability report, and no Scope 1/2/3 emissions disclosure as of March 2026.[1] The estimates in this document rely on Jegham et al. (2025)[2] as the sole empirical anchor for measured DeepSeek models (R1 and V3, January 2025 vintage). Current models (V3.1, V3.2) are estimated via throughput-based scaling.[3] DeepSeek operates inference on its own Chinese data centres and via Microsoft Azure,[4] creating a dual-infrastructure problem with dramatically different carbon outcomes. The Chinese grid carbon intensity (~0.6 kgCO_2e/kWh)[5] is approximately double that of US providers, making DeepSeek's own-server inference the most carbon-intensive of any major provider. Because the routing split between pathways is undisclosed, no single provider-wide figure can be defensibly stated. This paper presents three scenario pathways, not a single default. Confidence ranges from LOW to VERY LOW across all estimates. All figures should be read as scenario-based modelled ranges, not measured values. We invite DeepSeek to provide verified data.*

Scope notice: *This document covers operational inference emissions only. It excludes training, embodied hardware emissions, water consumption, and end-of-life disposal. Estimates represent grams of CO_2 equivalent (gCO_2e) per 1,000 tokens of model output, using location-based Scope 2 accounting as the primary metric.[6] CEAF (Clean-Energy Adjustment Factor)-adjusted figures are provided as a supplementary decision-support metric, not for offsetting calculations. These estimates do not constitute a company-wide environmental accounting for DeepSeek and should not be represented as such.*

Critical scope caveat: *DeepSeek's models are widely self-hosted by third parties on diverse hardware and grids worldwide. The carbon intensity of self-hosted inference depends entirely on the hosting party's infrastructure and is outside the scope of this analysis. The estimates below apply only to inference via DeepSeek's own API and via Azure-hosted endpoints.*

1. Executive Summary

DeepSeek presents the starkest infrastructure-driven carbon story in the InferenceCarbon reference series. While the models themselves — particularly DeepSeek-V3's Mixture-of-Experts (MoE) architecture with 671B total but only 37B active parameters[7] — represent genuine efficiency innovations, the carbon intensity of inference is dominated by two factors that overwhelm architectural gains: the Chinese electricity grid's carbon intensity (~0.6 kgCO_2e/kWh, roughly double the US average)[8] and the absence of any clean energy disclosure or procurement.

Jegham et al. (2025) measured both DeepSeek-R1 and DeepSeek-V3 on two infrastructure pathways: DeepSeek's own Chinese servers (H800 GPUs, PUE (Power Usage Effectiveness) 1.27, CIF (Carbon

[1]Earth911 (March 2026). "Your AI Carbon Footprint: What Every Query Really Costs." https://earth911.com/business-policy/your-ai-carbon-footprint-what-every-query-really-costs/
[2]Jegham, N. et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference." arXiv:2505.09598v6, November 2025. https://arxiv.org/abs/2505.09598
[3]Artificial Analysis (2026). LLM Leaderboard and model benchmarks. https://artificialanalysis.ai/models/
[4]Microsoft Learn (2026). "Get started with DeepSeek-R1 in Azure AI Foundry Models." https://learn.microsoft.com/en-us/azure/ai-foundry/foundry-models/tutorials/get-started-deepseek-r1
[5]Ember (2024). Global Electricity Review 2024 — China carbon intensity 581 gCO_2/kWh (2023 data). https://ember-energy.org/latest-insights/global-electricity-review-2024/
[6]GHG Protocol (2025). Scope 2 Guidance — upcoming revisions. https://ghgprotocol.org/blog/upcoming-scope-2-public-consultation-overview-revisions
[7]DeepSeek-AI (2024). "DeepSeek-V3 Technical Report." https://arxiv.org/abs/2412.19437

Intensivity Factor) 0.6) and Microsoft Azure (H200/H100 GPUs, PUE 1.12, CIF 0.35).[9][10] The results were dramatic: models on DeepSeek's own servers exhibited carbon emissions nearly six times higher than their Azure-hosted counterparts.[11] (At the time of writing, DeepSeek may be transitioning from H800 chips to Blackwell chips. This will affect gCO2e calculations, however the carbon intensivity of the Chinese energy grid remains a dominating factor. Also, we understand that DeepSeek is building an underwater data centre off Hainan island, and this may reduce PUE. We have no data on this.)

Because the routing split between these pathways is undisclosed, this paper does not designate a single "primary" provider-wide figure. Instead, following adversarial review, we present three scenario pathways throughout: Chinese-hosted, Azure-hosted, and a blended unknown-routing scenario. The infrastructure pathway is the dominant variable; model-version deltas are secondary.

External adversarial review characterised this paper as "the most justified location-based critique in the series" but warned that it "correctly sees that DeepSeek's carbon story is dominated by infrastructure, but then understates how much that same fact destabilises its model-level numbers." We accept this characterisation. All estimates should be read as scenario ranges, not model-level constants.

Summary of Scenario Estimates — DeepSeek-V3.2 (Non-Reasoning)

Scenario	Assumption	Range (gCO_2e/1k tokens, Estimated)	Central, Estimated	Confidence
A: Chinese DC	DeepSeek API; H800; CIF 0.60	3.50 – 9.00	5.86	Low
B: Azure-hosted	Azure AI Foundry; H200; CIF 0.35; CEAF 0.50	0.50 – 1.50 (location-based) *0.25 – 0.75 (CEAF-adjusted)*	0.91 (location-based) *0.46 (CEAF-adjusted)*	Medium-Low
C: Blended (unknown)	50/50 split (illustrative)	2.00 – 5.00	3.39	Very Low

Summary of Scenario Estimates — DeepSeek-R1 0528 (Reasoning)

Scenario	Assumption	Range (gCO_2e/1k tokens, Estimated)	Central, Estimated	Confidence
A: Chinese DC	DeepSeek API; H800; CIF 0.60	18.00 – 45.00	28.88	Very Low
B: Azure-hosted	Azure AI Foundry; H200; CIF 0.35; CEAF 0.50	1.00 – 4.00 (location-based) *0.50 – 2.00 (CEAF-adjusted)*	2.06 (location-based) *1.03 (CEAF-adjusted)*	Low
C: Blended (unknown)	50/50 split (illustrative)	10.00 – 25.00	15.47	Very Low

[9]Jegham et al. Table 1: DeepSeek (DS) infrastructure — DGX H800, PUE 1.27, WUE (site) 1.20 L/kWh, CIF 0.6 $kgCO_2e$/kWh.
[10]Jegham et al. Table 1: DeepSeek (Azure) infrastructure — DGX H200/H100, PUE 1.12, WUE (site) 0.30 L/kWh, CIF 0.35 $kgCO_2e$/kWh.
[11]Jegham et al. Section 5: "DeepSeek-R1 and DeepSeek-V3 deployed on DeepSeek's own servers exhibit water consumption and carbon emissions nearly six times higher than their Azure-hosted counterparts."

Note: Ranges incorporate anchor uncertainty, GEC (General Efficiency Correction, see below) range, throughput-proxy error, and CIF variability. Scenario B figures are shown at both location-based and CEAF-adjusted levels; CEAF 0.50 (Tier 1.5) applies to Microsoft Azure, reflecting its 100% annual renewable energy claim. For Scenario A (Chinese DC), CEAF = 0.00 (no clean energy disclosure), so location-based and CEAF-adjusted figures are identical. The blended scenario uses an illustrative 50/50 split of location-based figures; the actual routing share is unknown and may differ substantially. Reasoning-model ranges are especially wide because thinking-token ratios vary enormously by task. See Section 10 for the split uncertainty framework.

2. What Has Been Published

DeepSeek's disclosure position is comparable to Anthropic's — among the weakest of any major AI provider — but with the added complication of operating in a jurisdiction with no mandatory sustainability reporting framework for AI companies.

No energy data: DeepSeek has published no per-query energy consumption figure for any model. There is no equivalent to Altman's 0.34 Wh disclosure[12] or Google's measured 0.24 Wh median prompt figure.[13]

No sustainability report: DeepSeek has not published any environmental, sustainability, or ESG report. No Scope 1, 2, or 3 emissions have been disclosed.[14]

No clean energy disclosure: DeepSeek has made no public claim regarding renewable energy procurement, power purchase agreements, or carbon offsets for any of its data centres.

Open-source model weights: DeepSeek has published model weights (MIT license) for V3, R1, V3.1, and V3.2, plus detailed technical reports.[15][16] This is a significant transparency advantage for training efficiency analysis, but does not extend to infrastructure or operational sustainability disclosure.

Infrastructure signals: DeepSeek's primary compute is located in Chinese data centres, with major facilities in Inner Mongolia and an underwater data centre off Hainan Island.[17][18] The company uses NVIDIA H800 GPUs (the export-compliant variant for China).[19] DeepSeek models are also available via Microsoft Azure AI Foundry.[20]

Hardware uncertainty: Reuters reported in February 2026 that DeepSeek may have trained a recent model on NVIDIA Blackwell chips, linked to facilities in Inner Mongolia.[21] If confirmed, this would indicate DeepSeek's hardware story is shifting in ways this paper cannot cleanly capture. Any static "H800 = DeepSeek default" framing is therefore fragile.

3. Anchor Derivation

3.1 Source: Jegham et al. (2025)

[12]Altman, S. (June 2025). "The Gentle Singularity." https://blog.samaltman.com/the-gentle-singularity
[13]Google (August 2025). "Measuring the Environmental Impact of AI Inference." arXiv:2508.15734. https://arxiv.org/abs/2508.15734
[16]DeepSeek-AI (2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning." https://arxiv.org/abs/2501.12948
[17]PromptLayer (2025). "Where Are DeepSeek Data Centers Located." Inner Mongolia, Hainan underwater DC, Saudi Arabia. https://blog.promptlayer.com/where-are-deepseek-data-centers-located/
[18]South China Morning Post (February 2025). "China's subsea data centre could power 7,000 DeepSeek conversations a second." https://www.scmp.com/news/china/science/article/3299313/
[19]Jegham et al. Section 4.1: "DeepSeek, which operates under U.S. export restrictions, uses the H800, NVIDIA's export-compliant GPU for the Chinese market."
[21]Reuters (February 2026). "Exclusive: China's DeepSeek trained AI model on Nvidia's best chip despite US ban." https://www.reuters.com/world/china/chinas-deepseek-trained-ai-model-nvidias-best-chip-despite-us-ban-official-says-2026-02-24/

The anchor is drawn from Jegham et al. (arXiv:2505.09598v6),[22] the same infrastructure-aware benchmarking framework used across the InferenceCarbon series. Jegham measured DeepSeek-R1 and DeepSeek-V3 on both DeepSeek's own Chinese servers and Microsoft Azure, making this the only provider in the series with measured data across two distinct infrastructure pathways.

3.2 Infrastructure Specifications

Parameter	DeepSeek (own servers)	DeepSeek (Azure)
Hardware	DGX H800	DGX H200/H100
Critical Power (kW)	10.20	10.20
PUE	1.27	1.12
WUE (Water Usage Effectiveness) site (L/kWh)	1.20	0.30
WUE source (L/kWh)	6.016	4.35
CIF ($kgCO_2e$/kWh)	0.91	0.35

The CIF difference alone (0.60 vs 0.35) represents a 1.71× carbon multiplier. Combined with higher PUE (1.27 vs 1.12, a 1.13× multiplier) and less efficient hardware (H800 vs H200), the total carbon difference is approximately 5–6× between the two pathways.[23]

3.3 Measured Energy Data (Jegham et al.)

Jegham et al. measured energy consumption across three prompt configurations. The following figures are for DeepSeek's own Chinese server infrastructure:[24][25]

Model	Short (Wh)	Medium (Wh)	Long (Wh)
DeepSeek-V3 (DS)	2.777 ± 0.223	8.864 ± 0.724	13.162 ± 1.126
DeepSeek-R1 (DS)	19.251 ± 9.449	24.596 ± 9.4	29.078 ± 9.725
DeepSeek-V3 (Azure)	0.742 ± 0.125	2.165 ± 0.578	3.696 ± 0.221
DeepSeek-R1 (Azure)	2.353 ± 1.129	4.331 ± 1.695	7.410 ± 2.159

Short = 100 input + 300 output tokens; Medium = 1,000 + 1,000; Long = 10,000 + 1,500. All in watt-hours with standard deviation. DeepSeek-R1 figures include reasoning/thinking tokens.

Adversarial review caveat: These workload-normalised anchors apply to a particular measurement regime. They are not automatically a general per-1k-token truth across contexts with different input lengths, long-context memory pressure, reasoning overhead, or serving optimisations. MoE models can have unusual performance characteristics under different routing and context patterns. The tables look more universal than they should.

3.4 Deriving gCO_2e per 1,000 Tokens

We select DeepSeek-V3 (short query) as the primary non-reasoning anchor: the measured model most directly comparable to current V3.x successors.

[25]Jegham et al. Figure 3/Table 4: DeepSeek-R1 (DS) long prompt energy = 29.075 Wh.

Scenario A (Chinese DC):

Anchor query: 2.777 Wh for 400 tokens (100 input + 300 output). Energy per 1,000 tokens: 2.777 ÷ 400 × 1,000 = 6.943 Wh/1k tokens. Location-based carbon (CIF 0.60): 6.943 × 0.60 = 4.166 gCO_2e/1k tokens.

Scenario B (Azure):

Anchor query: 0.742 Wh for 400 tokens. Energy per 1,000 tokens: 1.855 Wh/1k tokens. Location-based carbon (CIF 0.35): 1.855 × 0.35 = 0.649 gCO_2e/1k tokens.

The 6.4× carbon difference between pathways (4.166 vs 0.649) for the same model underscores that infrastructure, not model architecture, is the dominant variable for DeepSeek's carbon intensity.

4. Scaling to Current Models

4.1 Method: Inverse Throughput Ratio

Consistent with the series, we use the inverse throughput ratio as the primary scaling proxy from the DeepSeek-V3 anchor to current models (V3.1, V3.2). Formula: Model estimate = Anchor energy × (Anchor TPS (Tokens per Second) ÷ Model TPS) × GEC (Generational Efficiency Correction).

Adversarial review caveat: This formula collapses multiple unknowns into one throughput-based scaling rule. A model's visible throughput can shift due to batch size, queueing policy, speculative decoding, KV-cache behaviour, MoE expert routing, hidden reasoning-token generation, chip architecture, and latency targets. Once routing ambiguity and hardware-shift uncertainty are admitted, TPS is no longer a clean proxy for joules per token. The scaling estimates that follow lean harder on the throughput proxy than the evidence robustly supports.

4.2 Throughput Data

Model	Output TPS	Source	Mode
DeepSeek-V3 (anchor)	~38	Approximate (no first-party AA data available)	Non-reasoning
DeepSeek-V3.2	24.3	Artificial Analysis (DeepSeek API)	Non-reasoning
DeepSeek-V3.2 Exp	37.7	Artificial Analysis (DeepSeek API)	Non-reasoning
DeepSeek-V3.1	~45	Approximate (no first-party AA data available)	Non-reasoning
DeepSeek-R1 0528	~15	Approximate (no first-party AA data available)	Reasoning

Sources: DeepSeek V3.2[26]; V3.2 Exp[27]; V3.1[28]; R1 0528.[29]

4.3 Generational Efficiency Correction (GEC)

For DeepSeek V3.x models relative to the V3 anchor: GEC range 0.85–0.95, midpoint 0.90. The narrower GEC range reflects that V3.1 and V3.2 are incremental updates to the same MoE architecture

[26]Artificial Analysis: DeepSeek V3.2 (Non-reasoning). 24.1 TPS, Intelligence Index 32. https://artificialanalysis.ai/models/deepseek-v3-2
[27]Artificial Analysis: DeepSeek V3.2 Exp. 37.7 TPS, released September 2025. https://artificialanalysis.ai/models/deepseek-v3-2-0925
[28]Artificial Analysis: DeepSeek V3.1. Intelligence Index 28. https://artificialanalysis.ai/models/deepseek-v3-1
[29]Artificial Analysis: DeepSeek R1 0528. Intelligence Index 27. https://artificialanalysis.ai/models/deepseek-r1

(685B total, 37B active), not a new model generation.[30] For DeepSeek-R1: No GEC applied. R1 is built on the V3 base and the Jegham measurement is direct.

5. The Dual-Infrastructure Problem

DeepSeek is unique among major providers in offering inference through two radically different infrastructure pathways: its own Chinese data centres (via the DeepSeek API at api.deepseek.com) and Microsoft Azure (via Azure AI Foundry or third-party providers hosting open-weight models).[31] The carbon difference between these pathways is the largest of any provider in this series.

5.1 Why We Do Not Designate a Single Primary Figure

Following adversarial review, this paper does not designate a single "primary" provider-wide number. The routing split between Chinese DC and Azure is undisclosed. A default can be a scenario choice, not a provider-wide fact. Presenting one privileged number would imply knowledge of traffic routing that we do not possess.

Instead, every estimate in this paper is explicitly labelled as one of three scenarios: (A) Chinese self-hosted, (B) Azure-hosted, or (C) blended unknown-routing. The blended scenario uses an illustrative 50/50 split; the actual ratio could range from 90/10 to 10/90.

5.2 Three-Scenario Comparison (DeepSeek-V3.2, non-reasoning)

	Scenario A: Chinese DC	Scenario B: Azure	Scenario C: Blended 50/50
Est. Wh/short query	3.91	1.04	2.48
CIF ($kgCO_2e/kWh$)	0.60	0.35	~0.48
gCO_2e/1k tokens (central)	~5.86	~0.91	~3.39
Carbon ratio vs Azure	6.4×	1.0×	3.7×

The total carbon difference compounds two independent factors: (A) higher energy per query on H800 hardware with higher PUE, and (B) a CIF that is nearly double. These are separated in the uncertainty framework (Section 10).

6. Per-Model Estimates (Non-Reasoning)

All estimates are presented as three-scenario ranges. Ranges incorporate anchor uncertainty, GEC range, throughput-proxy error, and CIF variability. These are scenario summaries, not model-level constants.

6.1 DeepSeek-V3.2 (Non-reasoning)

DeepSeek's current flagship general-purpose model, released December 2025.[32] 685B total parameters, 37B active (MoE). Intelligence Index 32. Output speed 24.1 TPS (DeepSeek API).

Derivation: 2.777 × (38 ÷ 24.3) × 0.90 = 3.908 Wh per short query (Scenario A)

[30]Sebastian Raschka (December 2025). "A Technical Tour of the DeepSeek Models from V3 to V3.2." https://magazine.sebastianraschka.com/p/technical-deepseek

Scenario	Range (gCO_2e/1k tokens, Estimated)	Central, Estimated	Confidence
A: Chinese DC	3.50 – 9.00	5.86	Low
B: Azure	0.50 – 1.50	0.91	Medium-Low
C: Blended 50/50	2.00 – 5.00	3.39	Very Low

6.2 DeepSeek-V3.1

Released August 2025. 685B total, 37B active. Intelligence Index 28.[33]

Derivation: 2.777 × (38 ÷ 45) × 0.90 = 2.111 Wh per short query (Scenario A)

Scenario	Range (gCO_2e/1k tokens, Estimated)	Central, Estimated	Confidence
A: Chinese DC	2.00 – 5.00	3.17	Low
B: Azure	0.30 – 0.80	0.49	Medium-Low
C: Blended 50/50	1.20 – 2.90	1.83	Very Low

7. Reasoning Models

Reasoning-model estimates are the least reliable in this paper. The hidden thinking-token overhead varies enormously by task — from near-zero for simple queries to 10–50× the visible output for complex mathematical or coding tasks. The following estimates exist on a continuous spectrum and should be treated as scenario summaries over a task distribution the paper cannot observe, not as model-level constants.

7.1 DeepSeek-R1 0528

Updated reasoning model, released May 2025.[34] Built on V3 base with reinforcement-learning-derived chain-of-thought reasoning. Intelligence Index 27 (heavily reasoning-weighted). Extremely slow throughput (~15 TPS) due to extended thinking token generation. Direct Jegham anchor: 19.251 Wh per short query (DS servers). No GEC applied.

Scenario	Range (gCO_2e/1k tokens)	Central	Confidence
A: Chinese DC	18.00 – 45.00	28.88	Very Low
B: Azure	1.00 – 4.00	2.06	Low
C: Blended 50/50	10.00 – 25.00	15.47	Very Low

7.2 DeepSeek-V3.2 (Reasoning Mode) — Illustrative Only

V3.2 includes an integrated reasoning mode that generates thinking tokens like R1 within the V3.2 inference framework.[35] Following adversarial review, we demote this estimate to illustrative status. There is no direct measurement. The figure is an interpolation between non-reasoning V3.2 and measured R1 behaviour — a heuristic placeholder, not a proper estimate.

Illustrative range (Scenario A, Chinese DC): 8.00 – 25.00 gCO_2e/1k tokens. This figure should not be used for comparison or display purposes without explicit qualification.

8. Three-Layer Disclosure Structure

Layer	Metric	V3.2 (Scenario A)	V3.2 (Scenario B)	R1 (Scenario A)
Layer 1: Energy	Wh per short query	3.91	1.04	19.25
Layer 2: Location-based	gCO_2e/1k (range)	3.50–9.00	0.50–1.50	18.00–45.00
Layer 3: CEAF-adjusted	gCO_2e/1k	3.50–9.00	0.25–0.75	18.00–45.00

Note: For Scenario A (Chinese DC), Layers 2 and 3 are identical because CEAF = 0.00 (Tier 3 — no clean energy disclosure). This is the only provider in the series where CEAF adjustment provides zero benefit. For Scenario B (Azure), CEAF 0.50 (Tier 1.5) applies.

9. Clean Energy Adjustment Factor (CEAF)

Pathway	Annual RE	Hourly CFE	CEAF Tier	CEAF
DeepSeek (Chinese DC)	No disclosure	N/A	Tier 3	0.00
Microsoft Azure	100% (Feb 2026)	Not published	Tier 1.5	0.50

DeepSeek's Chinese data centres receive the lowest possible CEAF score. China's electricity grid remains heavily coal-dependent (~580 gCO_2/kWh national average),[36] and DeepSeek has made no public commitment to renewable energy procurement. The CEAF framework's purpose — to differentiate providers making genuine clean energy investments — here serves to highlight the complete absence of such investment.

10. Uncertainty Framework

Following adversarial review, we present the dominant uncertainty sources in two distinct categories. These uncertainties are independent and compound. The missing provider disclosure is not a minor limitation; it is the defining condition of this paper.

Category	Source	Direction	Approx. magnitude	Resolvable by
A: Energy	Jegham Monte Carlo (anchor)	Known (±7% SD)	Small	Replication studies
A: Energy	GEC placeholder (V3→V3.x)	Unknown	Moderate (±10%)	New measurements
A: Energy	Throughput proxy confounders	Unknown	Moderate to large	Direct energy measurement
A: Energy	Token composition (short-query mix)	Unknown	Moderate	Usage-mix disclosure
A: Energy	H800 firmware limitations	Unknown direction	Potentially moderate	Hardware disclosure
B: Carbon	Chinese CIF regional variation	Known range (0.50–0.65)	Moderate (±15%)	Provider location disclosure
B: Carbon	Hardware pathway (H800 vs H200)	Known direction (H800 worse)	Large (1.5–2×)	Provider routing disclosure
B: Carbon	Routing split (Chinese/Azure)	Unknown	Dominant (≤5–6×)	Provider routing disclosure
B: Carbon	Blackwell chip transition	Unknown direction	Potentially large	Hardware disclosure

Category A uncertainties affect the energy estimate (Wh per query). Category B uncertainties affect the carbon conversion (gCO_2e per Wh). The routing-split uncertainty (B) is the single largest uncertainty in this paper and cannot be resolved without provider disclosure from DeepSeek. It alone can produce a 5–6× variation in the final carbon figure for the same model and same query.

11. Why Location-Based Accounting Should Be Preferred

The argument for location-based Scope 2 accounting is at its strongest for DeepSeek. There is no market-based accounting to contrast with because there are no renewable energy claims to evaluate. The Chinese grid's carbon intensity is a physical fact that directly determines the carbon content of every kilowatt-hour consumed during inference. With CEAF at zero for Chinese DC infrastructure, location-based is the only defensible metric.

12. Cross-Validation

Against Jegham/Altman: Jegham's GPT-4o short-query figure (0.423 Wh) aligns within 19% of Altman's 0.34 Wh claim,[37] giving confidence in the framework's calibration.

[37]Jegham et al. Section 5.3: Validation — Jegham GPT-4o short query (0.423 Wh) aligns within 19% of Altman 0.34 Wh claim.

Against Mistral LCA: Jegham's Mistral Large 2 estimate aligned within one standard deviation of Mistral's own published LCA figure,[38] further validating the methodology.

Against eco-efficiency ranking: Jegham's cross-efficiency DEA ranked DeepSeek-R1 (DS) at 0.058 and DeepSeek-V3 (DS) at 0.060 — the lowest scores of all 30 models measured.[39] Despite competitive intelligence ratings, the combination of high energy consumption and high CIF produces the worst eco-efficiency in the benchmark.

Against infrastructure comparison: The ~6× carbon difference between DeepSeek's own servers and Azure[40] is internally consistent: CIF ratio (1.7×) × PUE and hardware efficiency differences (~3×) ≈ 5–6× total.

13. Confidence Assessment

Estimate	Confidence	Dominant uncertainty
Anchor (V3, DS)	Medium	Single source; Chinese DC infrastructure assumed
Anchor (V3, Azure)	Medium	Single source; standard Azure CIF
V3.2 (non-reasoning)	Low	Throughput proxy; routing unknown
V3.1	Low	Third-party throughput; older version
R1 0528 (reasoning)	Very Low	Token ratio highly variable by task
V3.2 (reasoning)	Very Low	No measurement; heuristic interpolation only
Blended scenario	Very Low	Routing split unknown; 50/50 is illustrative
CEAF (Chinese DC)	High	Tier 3 = 0.00, no ambiguity

14. Limitations

1. Single empirical source. All anchor data comes from Jegham et al. (2025). No second independent measurement exists.

2. No provider disclosure. DeepSeek has published no energy, emissions, or sustainability data whatsoever. This is not a minor limitation; it is the defining condition of this paper.

3. Measured models are superseded. Jegham measured V3 and R1 (January 2025 vintage). Current models (V3.1, V3.2) are estimated via throughput scaling, which is a noisy proxy confounded by batching, KV-cache, MoE routing, and latency targets.

4. Chinese grid CIF uncertainty. We use 0.6 $kgCO_2e/kWh$ per Jegham. IEA forecasts suggest 0.505–0.565 for 2024–2026.[41] Regional variation within China is substantial — Inner Mongolia (coal-heavy) may exceed the national average.

5. Reasoning token variability. R1 and V3.2 reasoning-mode energy consumption varies enormously by task complexity. Central estimates are scenario summaries over a task distribution the paper cannot observe.

6. Hardware in flux. Reuters reported in February 2026 that DeepSeek may have accessed Blackwell chips.[42] Any static "H800 = DeepSeek default" framing is fragile.

[38]Mistral AI (2025). "Our Contribution to a Global Environmental Standard for AI." https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai
[39]Jegham et al. Appendix C: Cross-efficiency DEA scores. DeepSeek-R1 (DS) = 0.058, DeepSeek-V3 (DS) = 0.060 — lowest of 30 models.
[41]IEA (2025). Electricity Mid-Year Update 2025: China forecast 565 gCO_2/kWh (2024), declining to 505 by 2026. https://www.iea.org/reports/electricity-mid-year-update-2025/emissions

7. Routing unknown. The split between Chinese DC and Azure inference is undisclosed. This is the single largest source of uncertainty, capable of producing a 5–6× variation in the final carbon figure.

8. MoE efficiency captured but not decomposed. DeepSeek's MoE architecture should provide genuine efficiency gains relative to dense models, but this is already embedded in the Jegham measurement. We cannot decompose the architectural contribution.

9. Operational inference only. Training, embodied hardware, water, and Scope 3 emissions excluded.

10. Third-party hosting not covered. DeepSeek models are widely self-hosted. The carbon intensity depends on the hosting party's infrastructure.

11. Not a company-wide accounting. These estimates cover inference-phase operational emissions for specific model-API pathways. They should not be represented as DeepSeek's overall environmental footprint.

15. Response to Adversarial Review

An external adversarial review was conducted by GPT-5.4 Thinking (OpenAI) on 13 March 2026, using dual personas: a PhD technologist specialising in data-centre technologies, and a senior sustainability professional. The review was conducted on v1.0 of this paper.

Summary verdict: "This paper correctly sees that DeepSeek's carbon story is dominated by infrastructure, but then understates how much that same fact destabilises its model-level numbers." The technologist called it "a solid scenario-analysis memo, not a measurement-grade carbon model." The sustainability reviewer called it "the most justified location-based critique in the series, but still not safe enough for clean public model-by-model comparisons."

We accept these characterisations.

Three changes insisted upon and accepted:

1. Remove single primary provider number. Restructured entire paper around three explicit scenarios (Chinese DC, Azure, Blended unknown) rather than designating one "primary" pathway. Every table now labels its scenario. No privileged default is presented unless routing shares are disclosed. Implemented throughout v1.2.

2. Make uncertainty display much firmer. Added Section 10 (Uncertainty Framework) explicitly separating Category A (energy/Wh) from Category B (carbon/CIF) uncertainties. Routing-split uncertainty identified as the single dominant uncertainty, capable of producing 5–6× variation. Missing provider disclosure elevated from a limitation to "the defining condition of this paper." Implemented throughout v1.2.

3. Demote V3.2 reasoning estimate. V3.2 reasoning-mode moved to Section 7.2 as "Illustrative Only." Removed from executive summary tables. Explicit note that it is a heuristic interpolation, not a proper estimate, and should not be used for comparison or display without qualification. Implemented in v1.2.

Additional changes: Reuters Blackwell chip report flagged as hardware uncertainty (Section 2, Section 10, Limitation 6). Scope caveat strengthened to guard against over-reading as company-wide environmental accounting (Scope notice, Limitation 11). MoE throughput-proxy limitations elevated to inline adversarial review caveats (Sections 3.3, 4.1). Azure pathway confirmed via Microsoft Learn documentation.[43]

16. Recommendation

For InferenceCarbon.ai's carbon transparency tool, we recommend displaying all three scenario pathways for DeepSeek models, making the routing gap visually unavoidable. No single privileged "provider number" should be shown unless DeepSeek discloses actual routing shares:

Model	Scenario A: Chinese DC	Scenario B: Azure	Scenario C: Blended
V3.2 (non-reas.)	3.50 – 9.00 (c: 5.86)	0.50 – 1.50 (c: 0.91)	2.00 – 5.00 (c: 3.39)
V3.1	2.00 – 5.00 (c: 3.17)	0.30 – 0.80 (c: 0.49)	1.20 – 2.90 (c: 1.83)
R1 0528 (reas.)	18.00 – 45.00 (c: 28.88)	1.00 – 4.00 (c: 2.06)	10.00 – 25.00 (c: 15.47)

All scenarios should carry the label "Modelled estimate — no provider data." The uncertainty ranges should be visible in the display itself, not pushed into methodology notes. The routing gap (up to 5–6× between pathways) should be visually prominent.

V3.2 reasoning-mode should not appear in the main display. If shown at all, it should be in a secondary panel with explicit "illustrative only" qualification.

We commit to updating these estimates when: (a) DeepSeek publishes energy or sustainability data; (b) new independent measurements become available; (c) the Chinese grid CIF materially changes; (d) updated throughput data changes scaling ratios; or (e) the routing split between Chinese DC and Azure becomes known.

17. Key Sources

All sources are cited in footnotes throughout. Principal references:

Source	Used for
Jegham et al. (2025), arXiv:2505.09598v6	All anchor data — DeepSeek-V3, R1 energy (Chinese DC and Azure)
Artificial Analysis (2026)	Throughput for all DeepSeek models
Ember Global Electricity Review (2024)	Chinese grid carbon intensity (581 gCO_2/kWh)
IEA Electricity Mid-Year Update (2025)	Chinese grid CIF forecast (565→505 gCO_2/kWh)
Earth911 (March 2026)	No DeepSeek emissions disclosure confirmed
DeepSeek-V3 Technical Report (2024)	Architecture details (MoE, 671B/37B)
DeepSeek-R1 Paper (2025)	R1 training methodology, RL approach
Reuters (February 2026)	Blackwell chip access reports
Microsoft Learn (2026)	Azure AI Foundry DeepSeek hosting confirmed
PromptLayer / SCMP (2025)	Chinese data centre locations
GHG Protocol Scope 2 (2025)	Location-based accounting methodology

End of document. Version 1.4, 19 March 2026.

Prepared by InferenceCarbon.ai Reference Team