Deterministic Provenance Verification with Checker-in-the-Loop Repair: Small and Budget Language Models Reach Frontier Parity on Sourced Long-Form Generation

Bradley Norton
Norton Design Lab, LLC
bradley.norton@nortondesignlab.com

July 2026

Abstract

Large language models fabricate statistics when generating long-form content from source material, and LLM-based judges systematically fail to detect this behavior. We present a deterministic verification instrument that (i) generates synthetic evidence packs with seeded statistics, making claim-level ground truth available by construction; (ii) classifies every numeric claim in model output into provenance classes; (iii) applies a binary deployment gate; and (iv) feeds itemized violations back to the generating model for bounded repair. Across five evaluation runs (approximately 1,360 scored outputs, 20 models), we report four findings. First, raw prompting produces a 19× gap in gate pass rate between a frontier model (38/40, zero fabrications) and a locally served 35B model (2/40, 209 fabrications). Second, LLM judges from three model families rated 27.5% of gate-failing outputs at or above 4.5/5; a rubric intervention shifted two judge families by comparable magnitudes, indicating a shared ceiling on fabrication detection. Third, two bounded repair rounds driven by the deterministic checker eliminated the invented-statistic class in every model family tested, including the worst observed fabricator; three of four locally served 26–31B candidates reached an instrument-defined parity criterion against the frontier reference at zero marginal inference cost, while one 30B nano-class model plateaued below parity on structural and tagging residuals. Fourth, scaffolding efficiency does not transfer across model families: a strict compliance prompt that minimized cost for local models was the most expensive configuration for a reasoning-billed 744B mixture-of-experts model, tripling hidden reasoning tokens. We additionally document a reasoning-channel exhaustion failure mode that a provenance score alone rates as perfect and only a structural gate rejects. The verification layer requires no inference tokens; a reference implementation is released under the MIT license.

1 Introduction

Long-form generation from source material is a common production use of large language models (LLMs): reports, training materials, and marketing or research content that must restate statistics from an evidence base without alteration or invention. Two obstacles limit the substitution of smaller or locally served models for frontier models in this setting. The first is fabrication: models insert numeric claims that do not appear in the source, formatted indistinguishably from sourced claims. The second is evaluation: the prevailing quality-control mechanism, an LLM acting as a judge, exhibits reliability gaps on long-form output [1] and, as we show, systematically approves fabricated content.

A third, economic obstacle motivates the substitution in the first place. Production workloads such as retrieval-corpus enrichment consume tokens at rates that make per-token pricing a structural cost: a single contextual-enrichment pass over one document corpus in our production environment consumed approximately 5×10^8 tokens in one session, corresponding to roughly $1,900 at the observed blended commercial API rate ($3.79 per million tokens) and under $1 in marginal electricity cost on owned hardware.[1] Local inference is therefore economically forced for token-intensive workloads, and the question becomes whether its output can be trusted.

This paper argues that the trust question is resolved not by model selection but by a deterministic verification layer. We make five contributions:

1. **An instrument** combining parametric synthetic evidence packs (ground truth by construction, memorization excluded), deterministic claim-level provenance classification, a binary deployment gate, and a bounded checker-in-the-loop repair protocol, with first-draft and final scores reported separately so that scaffolding lift is measured rather than assumed (Section 3).
2. **A measurement of the raw capability gap**: 19× in gate pass rate between frontier and local models under identical prompting, with 209 fabricated statistics from the local model across 40 briefs (Section 4.1).
3. **Evidence that fabrication approval is a class property of LLM judges**: 27.5% of gate-failing outputs were judge-approved at 4.5+/5, and a rubric patch shifted judges from two independent model families by comparable magnitudes (Section 4.2).
4. **Conversion results**: bounded deterministic repair brought three locally served 26–31B models and one budget cloud 744B model to an instrument-defined parity criterion against a frontier reference, including full purging of fabrications from the worst observed fabricator (Sections 4.3–4.4).
5. **Two negative results with operational consequences**: scaffolding strategy rankings invert across model families under reasoning-token billing, and a reasoning-channel exhaustion failure mode produces empty or truncated outputs that a provenance score alone rates as perfect (Sections 4.5–4.6).

2 Related Work

Deterministic verification of LLM output. Two recent systems apply deterministic gates to a single pipeline's own output. AutoResearchClaw [5] maintains a whitelist registry of experimentally measured values and re-extracts every numeric claim from generated research papers, rejecting documents whose strict-section claims lack registry matches; its ablation shows that removing verification raises apparent acceptance (3/10 to 5/10) while admitting papers containing unmeasured values, and its case analysis observes that numeric verification is necessary but not sufficient. Nam et al. [8] describe integrity gates for clinical manuscript preparation, gating every stage transition with halt-on-failure and preferring deterministic re-executable checks wherever they suffice, motivated by the observation that self-critique inherits the blind spots that produce confident fabrication. Both systems verify one pipeline's output as a production safeguard; neither benchmarks models against one another, evaluates locally served models, nor measures repair as an intervention with a first-draft/final split. The present work uses the same verification primitive for a comparative question.

[1] The session was halted mid-run by a preconfigured $50 API spending limit and completed on local inference; the token figure is an operational estimate accurate to order of magnitude.

LLM-as-judge reliability. Chen et al. [1] document a substantial reliability gap for LLM judges on long-form output, with rubrics and reference answers helpful but not consistently sufficient. Our results extend this line with a fabrication-specific measurement against deterministic ground truth and a cross-family rubric intervention. Judge selection in our second instrument followed published human-agreement evidence [10].

Hallucination detection and self-correction. SelfCheckGPT [6] detects hallucination without ground truth via sampling consistency; FActScore [7] decomposes long-form output into atomic claims scored against a knowledge source. Our setting differs in that seeded synthetic packs make ground truth available by construction rather than estimation, permitting exact claim-level classification. The repair protocol is grounded in the finding that LLMs cannot reliably self-correct reasoning without external feedback [3]; the checker supplies that feedback deterministically. Naser [9] quantifies reference fabrication in AI-assisted academic writing across models, a bibliographic analogue of the statistical fabrication measured here.

Efficiency rebound. The economic discussion in Section 5 draws on the rebound-effect literature: the original backfire claim of Jevons [4], its critical modern review [12], the Khazzoom–Brookes formulation [11], and the long-run lighting consumption series of Fouquet and Pearson [2].

3 Instrument and Method

3.1 Synthetic evidence packs

Each brief pairs a topic prompt with an evidence pack of $k \in \{6, 12\}$ seeded statistics, generated parametrically from a seed tag so that (i) every numeric value's provenance is known by construction and (ii) no pack can have been memorized. Packs optionally include *collision* values, near-identical numbers attributed to different sources, as a difficulty pivot. The generation contract requires a fixed section structure, a self-audit footer, a minimum number of distinct pack statistics, and a source tag (`[S`n`]`) adjacent to every numeric claim.

3.2 Provenance classification and deterministic score

Let y be a model output and P the set of canonicalized pack values. A deterministic extractor produces the claim set $C(y)$ of numeric claims in y (digit forms, currency, percentages; word-form numerals in the production extractor). Each claim c receives a class:

$$\ell(c) \in \{\mathsf{tagged\_pack}, \mathsf{never\_tagged}, \mathsf{verbalized}, \mathsf{rounded}, \mathsf{invented}, \mathsf{invented\_tagged}\},$$

where tagged_pack denotes an exact pack value with an adjacent source tag (the only non-violating class); never_tagged an exact pack value lacking a tag; rounded/verbalized approximations or word-form restatements of pack values; invented a value absent from P; and invented_tagged an absent value carrying a fabricated source tag. The deterministic score is

$$\mathrm{det}(y) \;=\; \max\Big(0,\; 5 - \sum_{c \in C(y)} w\big(\ell(c)\big)\Big),$$

with deduction weights w increasing in severity (0.5 for never_tagged and rounded; 0.75 for invented; 1.0 for invented_tagged). The binary gate is

$$G(y) = 1 \iff \text{no violating claims} \wedge \text{footer present} \wedge \text{all sections present} \wedge |\{\text{distinct tagged pack values}\}| \geq m.$$

A single classification function serves both the repair loop and the analyzer, eliminating divergent read paths. We denote by det@0 the score of the first draft prior to any repair, and by det the final score; the difference is the measured scaffolding lift.

3.3 Checker-in-the-loop repair

On gate failure, the checker's itemized violations (claim text, class, and prescribed correction) are returned verbatim to the generating model together with the evidence pack, for at most R rounds ($R = 2$ throughout). No LLM judge participates at any point in this instrument. Four prompting strategies were crossed with repair: BASELINE (brief verbatim), STRICT (hardened compliance system message), FEWSHOT (brief plus one gate-verified exemplar), and PLAN_WRITE (statistic-allocation plan followed by generation; two calls).

3.4 Parity criterion and reference

A cell (model $\times$ strategy $\times$ repair budget, $n = 10$ briefs) is at *parity* when its gate rate is ≥ 0.95 and mean final det ≥ 4.85 relative to the best reference cell. The reference is GPT-5.6 under BASELINE with repair available; it recorded 70/70 gate passes at det 5.00 across all seven cells of the second run and reproduced this bar exactly (10/10, det 5.00, zero repair rounds) when re-run on byte-identical fixtures in the third run.

3.5 Experimental setup and integrity controls

Results derive from five runs: a blinded A/B replacement test (3 briefs $\times$ 10 models, deterministic columns plus blinded human quality scoring); two runs of a training-brief instrument with dual cross-family LLM judging layered above the deterministic checker (240 and 640 outputs; 16 models in the latter); and two runs of the long-form instrument described above (420 and 60 cells). Fixtures are SHA-256 pinned; the third long-form run inherited the second run's briefs by verified byte-identical copy. All monetary figures are billed API costs including hidden reasoning tokens, which are material: price sheets ignoring reasoning tokens understated two models' costs by 38–54% in the A/B test. Every number reported here regenerates programmatically from retained raw archives (outputs, metadata sidecars with billed costs, fixture hashes).

Three instrument defects were detected and corrected during the program by internal cross-checks: a silently dropped `invented` detection class (exposed by a smoke test whose gate rate moved from 2/3 to 0/3 after restoration, with 13 fabrications newly detected); a one-line tag-window search error that misclassified 20 correctly tagged claims (exposed by manual audit; fabrication detection verified unchanged pre- and post-fix at 87 detections); and a fixture-drift incident in which briefs were regenerated between judging passes, invalidating 234 of 640 judgments until three independent forensic methods confirmed the drift and hash-pinning was made load-bearing. We report these as evidence of the audit properties of deterministic layers rather than as threats: each defect was findable by reading code or comparing hashes, a property LLM judges do not share.

4 Results

4.1 The raw capability gap

Table 1 reports the complete 16-lane results of the 640-output run under identical raw prompting (40 briefs per lane, post-restoration checker). The frontier reference (GPT-5.6 Sol) passed 38/40 with zero

Lane	Gate	det	Invented	$/passing output	Judge
Commercial cloud					
GPT-5.6 Sol	38/40	4.87	0	0.119	4.96
GPT-5.6 Luna	31/40	4.42	12	0.030	5.00
Gemini 3.1 Pro	22/40	3.85	51	0.203	4.67
Qwen 3.7 Max	20/40	3.65	61	0.075	4.98
Claude Opus 4.8	13/40	3.39	73	0.312	4.74
Grok 4.5	4/40	2.36	231	0.244	4.99
DeepSeek V4 Flash	3/40	1.24	363	0.011	4.42
Claude Sonnet 5	2/40	2.00	118	1.094	4.81
GLM-5.2	1/40	1.94	248	0.302	4.96
DeepSeek V4 Pro	1/40	1.52	393	0.668	4.55
Spark-class candidates (served via commercial endpoints)					
Gemma 4 26B-A4B	2/40	1.86	135	0.020	4.48
Qwen 3.5 122B-A10B	2/40	1.76	245	0.131	4.51
Qwen 3.6 35B-A3B	2/40	1.48	185	0.068	4.34
GPT-OSS 120B	0/40	1.49	402	—	4.13
Qwen 3.6 27B	0/40	1.01	185	—	4.35
Locally served					
Qwen 3.6 35B-A3B (Q4, llama.cpp)	2/40	1.69	209	$\approx$0	4.95

Table 1: Complete run-three results of the training-brief instrument under raw prompting (no repair; $n = 40$ briefs per lane; 640 outputs). *Invented* counts invented-class claims only. *Judge* is the primary judge's (Kimi K2 Thinking) mean quality score on a 1–5 scale. $/passing output is billed cost including reasoning tokens; undefined for zero-pass lanes; the locally served lane incurs no marginal inference cost. Model identifiers and access dates are recorded in the run archive.

invented statistics (det 4.87); a locally served Qwen 3.6 35B model (4-bit quantization) passed 2/40 with 209 invented statistics (det 1.69), a 19$\times$ gap in gate rate. One budget cloud lane (GPT-5.6 Luna) achieved 31/40 at $0.0301 per passing output against a frontier lane (Claude Opus 4.8) at $0.3116. A high-fabrication cloud model collapsed on the collision axis (pass rate 0.60 to 0.25 as value confusability rose), demonstrating that difficulty must be treated as a surface rather than a point: models indistinguishable at low collision load separate under it.

Two columns of Table 1 warrant emphasis. First, the judge column: the primary judge's mean quality score exceeds 4.1/5 for every lane, including the two lanes that passed zero of forty briefs, and reaches 4.96 for a lane that passed one; this is the per-lane form of the divergence quantified in Section 4.2. Second, comparison with the preceding run shows that the detection-class restoration reordered the field: under the pre-restoration checker (run two of the training-brief instrument), Grok 4.5 and Claude Opus 4.8 tied at 38/40 and GLM-5.2 recorded 33/40; under the restored checker and refreshed fixture set, those lanes fell to 4/40 (231 invented), 13/40 (73 invented), and 1/40 (248 invented) respectively. Because the fixture set changed together with the checker, the reordering cannot be attributed to the detection fix alone, but the magnitude indicates that a large fraction of pre-fix passes contained undetected inventions. The frontier reference additionally recorded 57 off_pack_tagged claims (pack values attributed to an incorrect source, concentrated on the collision axis), a distinct attribution-error class from invention; its two gate failures contained zero invented claims.

Model	Cell	Gate	det	det@0	Invented
GPT-5.6 (reference)	BASELINE, R=0	10/10	5.00	5.00	0
	STRICT, R=2	10/10	5.00	5.00	0
gemma4-26b-a4b	BASELINE, R=0	7/10	4.80	4.80	2
	STRICT, R=2	10/10	5.00	5.00	0
gemma4-31b	BASELINE, R=0	4/10	4.25	4.25	6
	BASELINE, R=2	10/10	5.00	4.80	0
qwen3.6-27b	BASELINE, R=0	5/10	3.83	3.83	12
	STRICT, R=2	10/10	5.00	4.85	0
nemotron3-nano-30b	BASELINE, R=0	0/10	3.10	3.10	2
	FEWSHOT-RS, R=2	7/10	4.75	4.10	0
DeepSeek V4 Flash	BASELINE, R=0	0/10	1.02	1.02	23
	BASELINE, R=2	9/10	4.90	0.68	0

Table 2: Run-two condensed results ($n = 10$ briefs per cell): each model's raw cell and its best repair cell. The full 43-cell grid is available in the run archive. Invented counts are cell totals. The local models incur no marginal inference cost; nemotron3-nano-30b is the negative result discussed in the text.

4.2 Judge approval of fabricated content

In the 240-output run with dual judging, 66 outputs (27.5%) were rated $\geq 4.5/5$ by the LLM judge while failing the deterministic gate. Untagged fabricated statistics initially received the maximum pack-fidelity score from the primary judge; an explicit rubric patch moved the primary judge (Kimi K2 Thinking) from 5.00 to 4.73 on the affected dimension, and an independent judge from a second family (DeepSeek) to 4.60 on the same rows. A three-judge overlap analysis across three model families (Gemini 3.1 Pro, GPT-5.6, Kimi K2 Thinking) reproduced the gate-fail/judge-approve divergence, upgrading the finding from a single judge's blind spot to a shared property of LLM judges as a class. This is consistent with, and sharpens, the reliability gap reported by Chen et al. [1], and mirrors from the evaluation side the ablation of Liu et al. [5], in which removing deterministic verification admitted papers containing unmeasured values.

4.3 Deterministic repair converts every family tested

On the long-form instrument (run two: 420 cells across four locally served candidates, one high-fabrication cloud model, and the reference), bounded repair produced the results condensed in Table 2. A 31B local model (gemma4-31b) achieved 10/10 gate at det 5.00 under BASELINE with repair, i.e., with no exemplar or prompt engineering beyond the checker loop. Nine cells across three local models (gemma4-26b-a4b, gemma4-31b, qwen3.6-27b) met the parity criterion. The fourth local candidate, nemotron3-nano-30b, is a negative result: repair reduced its `invented` class to 0–1 per cell, but its gate rate plateaued at 2–7/10 on structural and tagging residuals across all five strategies, indicating that fabrication repair and contract compliance are separable capabilities and that not every small model converts.

The most efficient local configuration (gemma4-26b-a4b, STRICT) produced clean first drafts (det@0 5.00, 0.2 mean repair rounds) at 823 completion tokens per output versus 3,118 for the reference, a 3.8× token reduction for an identical verdict at zero marginal inference cost on owned hardware.

The worst observed fabricator (DeepSeek V4 Flash: det@0 1.02–1.20 raw; 16–23 fabrications per 10 briefs; 363 fabrications in the 40-brief run) converted to 9/10 gate at det ≈ 4.9 under both BASELINE and STRICT with repair, with every repaired brief ending at zero invented statistics; the residual violation class after repair was `never_tagged`, not `invented`. Fabrication, the failure mode of greatest practical

Cell	Gate	det	det@0	Fabrications	Mean rounds	$/passing output
GLM-5.2 BASELINE, R=0 (raw)	3/10	4.45	4.45	2	0.0	0.160
GLM-5.2 BASELINE, R=2	10/10	5.00	4.20	0	0.8	0.054
GLM-5.2 STRICT, R=2	10/10	5.00	5.00	0	0.6	0.100
GLM-5.2 FEWSHOT, R=2	10/10	5.00	4.60	0	0.4	0.058
GLM-5.2 PLAN_WRITE, R=2	10/10	5.00	4.20	0	0.6	0.047
GPT-5.6 reference, BASELINE, R=2	10/10	5.00	5.00	0	0.0	0.118

Table 3: Run-three results ($n = 10$ briefs per cell). All four scaffolded GLM-5.2 cells meet the parity criterion. The cheapest parity configuration costs $0.047 per verified output against the reference's $0.118 (a 60% reduction). Total billed cost of the 60-cell run: $4.25.

concern, was the most repairable violation class observed. One strategy–model interaction was negative: the FEWSHOT exemplar improved this model's first drafts but degraded repair convergence, leaving never_tagged residuals on 6/10 briefs.

4.4 Budget cloud confirmation and cost analysis

Run three evaluated GLM-5.2 (744B mixture-of-experts, approximately 40B active) across all four strategies plus a raw cell, against a same-fixture reference replication. Table 3 reports the full results; all monetary values are billed costs including reasoning tokens.

The model entered with a split prior: near-clean deterministic columns on the A/B instrument (zero untagged statistics) but 13 fabrications in a 3-brief smoke test on the training-brief instrument. On this instrument its raw cell produced 2 fabrications in 10 briefs, and every repair configuration purged them within two rounds. Fabrication in this model is real, low-rate, instrument-sensitive, and fully repairable under the protocol.

4.5 Strategy rankings do not transfer across families

The STRICT prompt, the efficiency-optimal configuration for the local gemma models, was the most expensive configuration for GLM-5.2 despite producing perfect first drafts. The mechanism is reasoning-token billing: STRICT induced 225,134 hidden reasoning tokens across the cell versus 75,823 for PLAN_WRITE (28,225 versus 11,844 total completion tokens; 574.6 s versus 195.7 s mean latency), placing STRICT at $0.100 per passing output, within 16% of the frontier reference's price. The cost-optimal path for a reasoning-billed model was the cheap draft plus deterministic repair ($0.047), not the compliance prompt. The FEWSHOT degradation observed for DeepSeek V4 Flash did not reproduce on GLM-5.2 (10/10 at the lowest round count, 0.4), indicating that the interaction is a property of the model rather than the exemplar. Both observations argue that scaffolding strategy must be measured per family rather than assumed to transfer.

4.6 Reasoning-channel exhaustion

Two of 52 GLM-5.2 generation calls (3.8%) failed by exhausting output in the hidden reasoning channel: one call emitted 3,906 tokens, all reasoning, and zero content; a resample of the same brief consumed 23,052 of a 24,000-token budget in reasoning and truncated approximately 950 tokens into the content. The empty output received det@0 = 5.00, a perfect provenance score, because zero extractable claims implies zero violations; the structural gate rejected it on missing sections. A provenance score alone

therefore rates an empty document as perfect: scoring layers that measure only what is present cannot penalize what is absent, and binary structural requirements are load-bearing. The same failure class was independently observed in our production ingestion pipeline (1,803 consecutive empty enrichment outputs caused by a chat-template reasoning-channel misconfiguration), where a deterministic preflight canary and a per-file completion-ratio gate detected it before any downstream cost was incurred. Operationally the failure is a per-endpoint reasoning-budget parameter issue rather than a model-quality issue.

5 Discussion

Routing implications. The results support a three-tier routing policy under a single verification layer: locally served models for token-intensive workloads (where the $19\times$ raw gap is closed by repair at zero marginal cost), budget cloud models where additional capability margin is desired (at 60% below frontier cost per verified output, with a reasoning-budget cap to preempt the exhaustion mode), and frontier models where blinded writing quality rather than provenance is the differentiator. The instrument measures the provenance and structure contract only; in the blinded A/B test, a frontier-class model won the quality scoring across three replications, and a gate-clean output can remain the worse piece of writing. The apparent contradiction between the local model's 2/40 raw result and its production use for retrieval-corpus enrichment resolves by task class: grounded summarization with the source stored verbatim confines fabrication risk to retrieval quality, and generation under the checker is a different result entirely.

Verification cost and rebound. The verification layer consumes no inference tokens; its cost is the domain understanding required to encode what constitutes a violation. Because it converts low-cost models to gate-clean output, its predictable aggregate effect is not reduced expenditure but expanded production, consistent with the rebound literature: efficiency gains in a general-purpose input expand consumption of its useful output [4, 11], with the largest effects for general-purpose technologies early in diffusion [12]; the seven-century lighting series of Fouquet and Pearson [2] (efficiency $\times 10^3$, consumption $\times 6.6 \times 10^3$) is the canonical illustration. Verified tokens are the useful output here; removing the trust constraint removes the final reason to ration them.

Auditability as the selection criterion. Deterministic layers share the blind-spot problem in principle: the dropped detection class and the tag-window error were checker defects. The relevant distinction is that these defects were findable and fixable by reading code and comparing hashes, whereas the judge ceiling on fabrication persisted across model families and rubric interventions. We therefore propose auditability, rather than accuracy alone, as the criterion for selecting which layer of an evaluation stack is made load-bearing.

6 Limitations

Cells are $n = 10$ briefs (long-form instrument) or $n = 40$ (training-brief instrument); consecutive ceiling results are strong evidence rather than proof, and confidence intervals on 10/10 cells are wide. The long-form fixture set derives from a single seed family. The strict-inversion result is one model on one endpoint. Fabrication counts on the cross-instrument comparison are small in both directions (2 in 10 briefs versus 13 in 3), supporting a directional claim of instrument sensitivity only. Latency and throughput are endpoint- and hardware-dependent and are not model properties. The simplified reference implementation omits the word-form and interposed-tag handling of the production extractor, where a

substantial fraction of engineering effort resides. The instrument does not measure writing quality, and parity claims are restricted to the provenance and structure contract.

7 Conclusion

A deterministic, token-free verification layer with bounded repair brought three of four locally served candidates and every cloud family tested to frontier-level performance on a claim-provenance contract, at 60% to effectively 100% cost reduction per verified output, while LLM judges from three families approved over a quarter of provably fabricated outputs. The layer additionally caught an empty-output failure mode that provenance scoring alone rates as perfect, and exposed a family-dependent inversion of scaffolding economics under reasoning-token billing. Across three instruments and twenty models, the pattern is consistent: the verification scaffold, not the generating model, is the load-bearing component for the provenance contract; the invented-statistic class was eliminated by repair in every family tested, and the informative differences between models lie in how they fail, with one nano-class model demonstrating that contract compliance and fabrication repair are separable capabilities.

Reproducibility. All tables regenerate programmatically from retained raw archives (outputs, billed-cost metadata, SHA-256 fixture pins). A reference implementation of the checker and repair loop (approximately 100 lines, standard library only) is released under the MIT license.

References

[1] Chen, J., Dong, Y., Li, H., Su, W., Zhou, Y., Zhang, M., Liu, Y., and Ai, Q. (2026). Benchmarking LLM-as-a-Judge for long-form output evaluation (LongJudgeBench). *arXiv:2606.01629*.

[2] Fouquet, R. and Pearson, P. (2006). Seven centuries of energy services: The price and use of light in the United Kingdom (1300–2000). *The Energy Journal*, 27(1):139–177.

[3] Huang, J., et al. (2023). Large language models cannot self-correct reasoning yet. *arXiv:2310.01798*.

[4] Jevons, W. S. (1865). *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-Mines*. Macmillan, London.

[5] Liu, J., Qiu, S., et al. (2026). AutoResearchClaw: Self-reinforcing autonomous research with human–AI collaboration. *arXiv:2605.20025*.

[6] Manakul, P., Liusie, A., and Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv:2303.08896*.

[7] Min, S., et al. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv:2305.14251*.

[8] Nam, Y., et al. (2026). Deterministic integrity gates for LLM-assisted clinical manuscript preparation: An auditable biomedical informatics architecture. *arXiv:2606.09500*.

[9] Naser, M. Z. (2026). How LLMs cite and why it matters: A cross-model audit of reference fabrication in AI-assisted academic writing and methods to detect phantom citations. *arXiv:2603.03299*.

[10] Pulipaka, S., Chen, O., Sharma, M., Bajwa, T. S., Raina, V., and Sheth, I. (2026). PersistBench: When should long-term memories be forgotten by LLMs? *arXiv:2602.01146*. ICML 2026.

[11] Saunders, H. D. (1992). The Khazzoom–Brookes postulate and neoclassical growth. *The Energy Journal*, 13(4):131–148.

[12] Sorrell, S. (2009). Jevons' Paradox revisited: The evidence for backfire from improved energy efficiency. *Energy Policy*, 37(4):1456–1469.