

GreenSwap: Verb Choice or Output Length? A Controlled Decomposition of Prompt-Level Energy Savings in LLM Inference

Ahsan Farabi, Israt Khandaker, Md. Abdul Ahad Minhaz,
Zannatul Zahan Meem, Ibrahim Khalil Shanto, Shahriar Mahmud

Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh

{afarabi221266, ikhandaker221263, mminhaz213072, zmeem213178, ishanto213193}@bscse.uui.ac.bd
shahriarmahmud@cse.uui.ac.bd

Abstract

Changing one word in a prompt can reduce large language model inference energy by up to 70%, but no prior work explains why. We introduce GreenSwap to find out. We run every prompt under three conditions: the original verb (C1), a swapped low-energy verb (C2), and the original verb with an explicit instruction to limit the response length (C3). Across three open-weight models, three task types, and 21,600 inferences, the short-answer instruction saves more energy and preserves higher quality than verb swapping in 7 of 9 settings. A token-energy correlation of $r \geq 0.998$ shows that inference energy depends only on output length and not on which verb was used. Verb swapping saves energy only because it sometimes makes the model write less, and on long-form generation it backfires by producing a completely different and lower-quality response. Scoped to English instruction-tuned 7–8B models, our results show that telling the model how long to respond is simpler, more reliable, and more energy-efficient than curating verb lists.

1 Introduction

Large language models are now widely deployed across industries, and their energy footprint during inference has become a growing environmental concern. While training costs have received significant public attention, inference dominates the total operational energy of deployed systems, since every user query triggers a full forward pass through billions of parameters. As model usage scales to millions of daily queries, even small per-query energy savings compound into meaningful reductions. This has motivated a search for lightweight, zero-modification strategies that reduce inference energy without altering model weights or architecture.

One promising direction is changing how prompts are worded. Recent work shows that swapping a single instruction verb, replacing “explain”

with “classify” for instance, can cut inference energy by up to 70% with no change to model weights (Adamska et al., 2025). This finding has attracted attention precisely because it requires no engineering effort. But it rests on an assumption that has never been tested: that the verb itself is responsible for the saving. An equally plausible explanation is that certain verbs simply happen to elicit shorter responses, and shorter responses naturally consume less energy (Samsi et al., 2023). If this second explanation is correct, then curating verb lists is unnecessary and practitioners could achieve the same or greater savings by telling the model directly how long its answer should be (Nayab et al., 2026). These two explanations are observationally equivalent in every prior study because none of them controlled for output length. No existing work directly compares verb swapping against an explicit length instruction at the per-inference level, which means the practical recommendation of curating verbs or controlling length has never been grounded in a controlled experiment. Our aim is not to challenge any single prior study, but to answer a broader scientific question that applies across this line of work: whether prompt-level energy savings arise from lexical choice itself or from the output length that a given choice happens to induce.

We introduce GreenSwap, a controlled three-condition experiment designed to separate these two explanations, as illustrated in Figure 1. For every combination of model, verb pair, and input item, we run three conditions. The first condition (C1) uses the original high-energy verb. The second condition (C2) replaces it with the low-energy alternative. The third condition (C3) keeps the original verb but appends an explicit instruction to respond within the same token budget that C2 naturally produced. If the verb carries an intrinsic effect, C2 should use less energy than C3. If output length is the sole driver, C3 should match

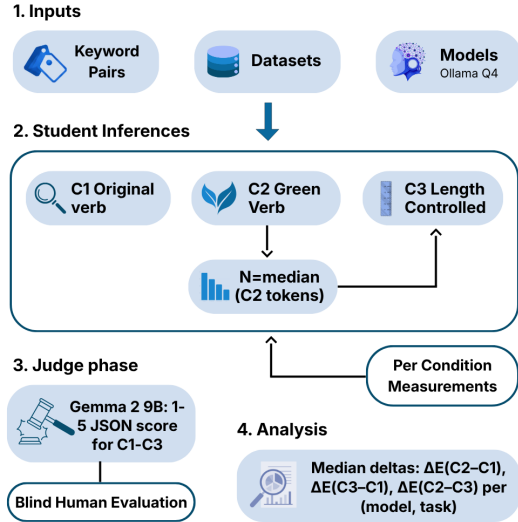


Figure 1: GreenSwap pipeline. Each combination of model, pair, and item runs C1, C2, and C3. C3’s budget N is derived at runtime from C2’s actual output length. Quality is assessed by a cross-family LLM judge, task-specific lexical metrics, and a blind human evaluation.

or outperform C2. This three-way design is the first to directly pit verb-level prompting against runtime length control across multiple models and task types.

Our main contributions are as follows.

- We propose GreenSwap, the first controlled decomposition of prompt-level energy savings that isolates lexical verb effects from output length effects through a matched three-condition design across three models and three task types, providing the first per-inference falsification of the lexical hypothesis.
- We show that a token–energy correlation of $r \geq 0.998$ across 21,600 inferences rules out any residual lexical mechanism, confirming that inference energy is a near-deterministic function of output length.
- We demonstrate that explicit length instructions (C3) are Pareto-dominant over verb swapping (C2) in 7 of 9 model–task settings, saving more energy while preserving or improving answer quality.
- We show that verb swapping degrades output quality on long-form generation tasks by producing semantically different responses rather

than shorter correct ones, confirmed by both an automated judge and a blind human evaluation ($\kappa = 0.83$).

- We introduce the compliance ratio as a practical diagnostic that explains model-dependent outcomes and establishes compliance auditing as a necessary step before deploying length-based interventions.

2 Related Work

The environmental cost of LLM inference has grown alongside model deployment at scale. Strubell et al. (2019) first quantified the energy cost of NLP training, and Luccioni et al. (2023) and Bannour et al. (2021) extended these measurements to inference, showing it dominates total operational energy in deployed systems. Samsi et al. (2023) further demonstrate that autoregressive token generation accounts for the vast majority of per-query energy, establishing output length as the primary axis along which inference cost can be reduced. Patterson et al. (2021) and Schwartz et al. (2020) situate these findings within a broader sustainability agenda for AI research, calling for lightweight strategies that reduce inference cost without sacrificing task performance.

Several strategies reduce inference cost without modifying model weights. Model quantization reduces memory and compute requirements at the cost of small quality degradation (Dettmers et al., 2022). Speculative decoding accelerates generation by using a smaller draft model to propose tokens that a larger model then verifies (Leviathan et al., 2023). Prompt compression reduces the number of input tokens processed per query (Jiang et al., 2023). Our work is orthogonal to all three: we target output length through prompt-level instructions rather than altering the model, the decoding algorithm, or the input. These approaches can be combined with GreenSwap for further savings.

Adamska et al. (2025) provide the most direct precedent for our work, systematically measuring per-verb energy across three tasks and three 7B models and reporting per-keyword energy rankings from which we construct our verb pairs. Rubel et al. (2025) complement this picture by showing that prompt engineering techniques targeting code generation can reduce the carbon emissions of Llama 3 on the CodeXGLUE benchmark. Both studies establish that prompt wording affects inference energy, but neither can determine whether this effect

is intrinsically lexical or simply a consequence of eliciting shorter outputs, as their designs cannot separate the two explanations. Our three-condition design is built specifically to falsify the lexical hypothesis.

Nayab et al. (2026) show that explicit length constraints in prompts reduce token generation while preserving or improving task accuracy, providing the motivation for our C3 condition. Han et al. (2025) similarly compress reasoning-chain length via prompt-level token budgets, but find that compression effectiveness depends critically on the chosen budget size, a dynamic we characterize directly through our compliance ratio. Poddar et al. (2025) benchmark response length across 12 decoder-only LLMs and find that models routinely generate responses far longer than necessary, motivating output-length control as a general efficiency lever. Li et al. (2024) propose generation directives that guide the autoregressive decoding process itself to reduce the carbon footprint of LLM inference, an alternative locus of control to the prompt-level interventions we study. Zhou et al. (2023) document that models comply with such instructions inconsistently across families and task types, a finding that directly foreshadows our compliance-ratio results. These lines of work are correlational and share a fundamental limitation: they cannot determine whether energy savings are driven by the prompt wording itself or by its effect on output length. GreenSwap is the first design to separate these accounts at the per-inference level.

We use a cross-family LLM judge to assess output quality across 21,600 inferences at a scale where human evaluation alone would be impractical. Zheng et al. (2023) introduce the LLM-as-judge paradigm and show it correlates well with human preferences on open-ended generation tasks. Wang et al. (2024) identify verbosity and conciseness biases in LLM judges, a concern directly relevant to our setting since our rubric structurally favors shorter outputs. We mitigate this through a cross-family judge, fixed JSON-structured prompts, lexical backup metrics, and a blind human evaluation, following best practices from both studies.

3 Methodology

3.1 Three-Condition Design

For each combination of student model m , verb pair p , and input item x , we generate three outputs under a fixed prompt template (Appendix C).

C1 (original verb). The prompt uses a high-energy verb v_o from Adamska et al. (2025): “ v_o the answer to the following question” for QA and generation, or “ v_o the sentiment as positive or negative” for sentiment.

C2 (green verb). Identical to C1 with v_o replaced by the low-energy alternative v_g . No other change is made.

C3 (length-instructed). Identical to C1 with the suffix “Answer in N tokens or fewer” appended, where $N = \max(5, \text{median}(T_o^{C2}))$; the floor of 5 prevents degenerate budgets on single-token tasks. C3 is a *length-requesting* condition, not a hard-cap baseline: compliance varies substantially across model families ($\bar{\rho} = 0.35\text{--}0.84$; Section 4.5), and the C2-vs-C3 ordering is fully predicted by this deviation (Section 5.4). We emphasize that C3 is not proposed here as a novel prompting technique; it serves as a controlled baseline that isolates the effect of output length from the effect of verb choice. Its purpose is diagnostic rather than prescriptive: if verb identity carried an intrinsic effect independent of length, C2 would be expected to outperform C3 on energy.

If verb identity has an intrinsic effect, C2 should use less energy than C3; if output length is the sole driver, C3 should match or outperform C2.

3.2 Verb Pairs, Datasets, and Models

We construct 24 verb pairs (8 per task) using the per-keyword energy rankings reported in Adamska et al. (2025) Table II (Appendix B). Three pairs appear in both QA and Generation, enabling cross-task replication (Section 5.3).

Datasets are three public benchmarks (50 validation items each): TriviaQA (Joshi et al., 2017) (rc.nocontext) for QA, ELI5 (Fan et al., 2019) via KILT (Petroni et al., 2021) for generation, and SST-2 (Socher et al., 2013) from GLUE (Wang et al., 2018) for sentiment. These represent low, high, and near-zero output-length headroom respectively. Total student inferences: $150 \times 8 \times 3 \times 2 \times 3 = 21,600$.

Student models are three instruction-tuned 7–8B models served via Ollama with Q4_K_M quantisation: qwen2.5:7b, llama3.1:8b, and mistral:7b-v0.3. All inferences use temperature=0, seed=42, num_ctx=2048, and num_predict=512. The judge is gemma2:9b-instruct-q4_K_M, chosen from

a different model family to reduce intra-family bias (Wang et al., 2024).

3.3 Measurement and Quality Assessment

Energy is measured with CodeCarbon (Lottick et al., 2019) in process mode at 0.5 s sampling (median of 2 runs per cell). We report $\Delta E_{X/Y} = (E_X - E_Y)/E_Y \times 100\%$; negative values indicate lower energy for X .

Quality uses three signals: (1) a Gemma 2 9B LLM judge on a 1–5 rubric; (2) lexical metrics (token F1 for QA, label accuracy for SST-2); and (3) a blind human evaluation on 100 generation outputs (Section 4.8).

Rubric limitation. The rubric structurally favors shorter outputs: correct+verbose=3, correct+concise=4. All C3 judge improvements therefore reflect conciseness *alongside* correctness. Lexical metrics and the human evaluation serve as rubric-free checks throughout.

3.4 Statistical Analysis

We compare conditions at the per-inference level using the Wilcoxon signed-rank test, a non-parametric test suited to paired energy measurements that do not assume normality. We use $n = 400$ paired observations per model–task stratum (50 items $\times$ 8 verb pairs) and apply Holm–Bonferroni correction across all 27 contrasts (9 model–task cells $\times$ 3 condition pairs) to control the family-wise error rate.

Effect sizes are reported as Cliff’s δ (Cliff, 1993), which measures the probability that a randomly drawn observation from one condition is larger than one from another. We follow the standard threshold of $|\delta| > 0.474$ for a large effect. We set $\alpha = 0.05$ throughout. Full test results with effect sizes for all 27 contrasts are in Appendix G.

4 Results

4.1 Headline Results

The pipeline ran without errors; truncation rate was 1.44% (C1 long-form outputs hitting num_predict=512). Table 1 presents per-(model, task) energy and quality results; Figure 2 shows bootstrap confidence intervals.

4.2 Task-Specific Pattern

QA. All three models save 41–58% energy with C2; judge scores are statistically unchanged ($p = 0.56$ – 0.63 , ns), and Token F1 is stable or improved

under C2 (Table 3), corroborating the result with a rubric-free metric. C3 saves more (58–81%) with judge gains of +0.28 to +0.53 (conciseness + correctness; see rubric note). QA is the only task where C2 constitutes a strict Pareto improvement.

Generation. C2 saves 29–67% energy but drops judge mean by -0.21 to -1.04 (all $p < 10^{-8}$). *Classify* and *translate* shift the *type* of response rather than its length: the model produces a topic label or paraphrase instead of an explanation (Appendix E). C3 saves 42–84% energy and preserves quality; blind human evaluation corroborates C3’s correctness advantage (Section 4.8).

Sentiment. Responses are already approximately one word, leaving no headroom for verb-driven compression. C2 saves only 2–6% with no quality change ($p = 0.37$ – 0.57 , ns); SST-2 accuracy is stable across all conditions (≥ 0.945 , Table 3), confirming no correctness loss.

4.3 Falsification Test: C2 vs. C3

Figure 3 shows per-cell $\Delta E_{C2/C3}$ distributions. For Qwen, C3 wins in 98–100% of cells because Qwen under-shoots its budget ($\tilde{\rho} = 0.35$). For Llama QA, the result is a statistical tie ($\delta = +0.03$) because Llama complies near-perfectly ($\tilde{\rho} = 0.95$) — precisely the signature of a length-mediated effect. For Llama Generation, C2 wins on energy in 77.5% of cells ($\delta = -0.55$, $\tilde{\rho} = 1.53$) but at a -1.04 judge-score cost confirmed by human raters, making it a Pareto loss. For Mistral, C3 wins in 65–90% of cells. Two contrasts are non-significant: $\Delta E_{C2/C1}$ on Qwen Sentiment ($p = 0.011$, $\delta = -0.10$) and $\Delta E_{C2/C3}$ on Llama QA ($\delta = +0.03$, tie).

4.4 Token–Energy Correlation

Pearson $r = 0.9996$ (Qwen), 0.9981 (Llama), 0.9996 (Mistral) across all 21,600 inferences (Figure 4), leaving less than 0.4% of energy variance for any residual lexical effect. This near-deterministic relationship is consistent with prior work (Samsi et al., 2023); our contribution is leveraging it to falsify the lexical hypothesis through the matched C2/C3 design. This near-deterministic relationship follows directly from how autoregressive decoding works: each additional output token requires one more forward pass through the full model, so total inference cost scales essentially linearly with the number of tokens generated, regardless of which verb triggered that generation.

Model	Task	Tok. (med.)		$\Delta E_{C2/C1}$	$\Delta E_{C3/C1}$	$\Delta E_{C2/C3}$	C3 wins (%) ^{††}	Judge (1-5) [†]		
		C1	C2					C1	C2	C3
Qwen 2.5 7B	QA	116	59	-47.5%*	-80.9%*	+181.6%*	98%	2.67	2.69	2.95*
Qwen 2.5 7B	Gen	374	130	-55.0%*	-84.0%*	+155.5%*	98%	4.16	3.13*	3.76*
Qwen 2.5 7B	Sent	61	57	-2.4% [§]	-69.2%*	+217.3%*	100%	3.49	3.51	3.95*
Llama 3.1 8B	QA	83	38	-58.2%*	-58.2%*	+3.3%*	51%	2.96	2.93	3.49*
Llama 3.1 8B	Gen	383	104	-66.6%*	-48.0%*	-34.5%*	22%	3.49	2.46*	3.72*
Llama 3.1 8B	Sent	64	60	-6.3%*	-53.0%*	+95.1%*	99%	3.69	3.68	3.88*
Mistral 7B	QA	110	63	-41.4%*	-66.5%*	+64.8%*	86%	2.59	2.62	2.90*
Mistral 7B	Gen	330	204	-29.4%*	-42.3%*	+15.1%*	65%	3.94	3.73*	3.88 [§]
Mistral 7B	Sent	57	49	-5.1%*	-38.5%*	+46.2%*	89%	3.50	3.59**	3.78*

Table 1: Per-(model, task) results ($n = 400$ per row). $\Delta E_{X/Y}$: median % energy change of X relative to Y ; * $p < 0.001$, ** $p < 0.01$, [§] $p < 0.05$ (Holm–Bonferroni). ^{††}**C3 wins (%)**: fraction of cells where C3 uses strictly less energy than C2. [†]Rubric favors conciseness (Section 3.3); bold C3 = significantly better than C1 ($p < 0.05$); C2* = significantly worse than C1.

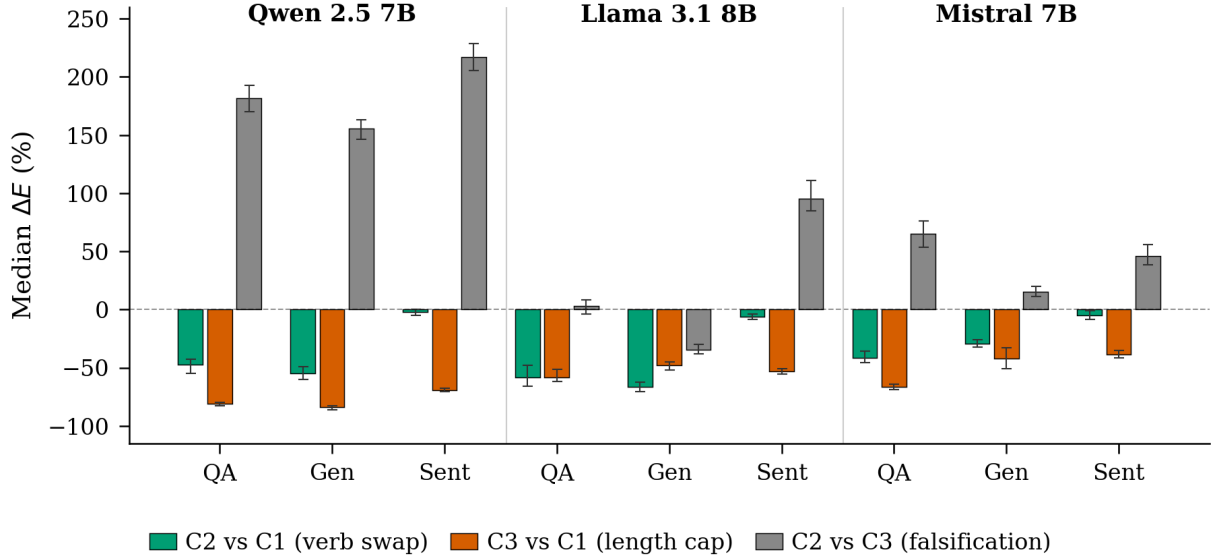


Figure 2: Median ΔE (%) per (model, task) with bootstrap 95% CIs ($n = 400, 500$ resamples). **Green**: C2 vs C1 (verb swap); **orange**: C3 vs C1 (length instruction); **gray**: C2 vs C3 (positive = C3 uses less energy; negative = C2 uses less energy).

4.5 Length-Budget Compliance

The compliance ratio $\rho = T_o^{(C3)}/N$ measures adherence to the prompted budget ($\rho = 1$ perfect; $\rho < 1$ under-shoot; $\rho > 1$ over-shoot). Table 2 and Figure 5 reveal three distinct regimes. Qwen severely under-shoots ($\tilde{\rho} = 0.35$), Llama overshoots on generation ($\tilde{\rho}_{\text{Gen}} = 1.53$), and Mistral is intermediate ($\tilde{\rho} = 0.73$). These regimes directly predict the C2-vs-C3 energy ordering (Section 5.4).

4.6 Answer Quality

Table 3 combines judge scores with lexical metrics. Token F1 values are low in absolute terms owing to the open-ended TriviaQA format (no retrieval context); relative differences between con-

Model	$\tilde{\rho}$	Mean	$\rho < 0.5$	$\rho > 1.0$
Qwen 2.5 7B	0.35	0.38	84.7%	0.7%
Llama 3.1 8B	0.84	1.16	23.6%	40.1%
Mistral 7B	0.73	0.74	21.8%	18.2%

Table 2: Length-budget compliance ($n = 1,200$ each). $\tilde{\rho}$: median actual / prompted tokens. Llama’s mean > 1 is driven by generation over-shoot ($\tilde{\rho}_{\text{Gen}} = 1.53$).

ditions are the relevant signal. C2 substantially degrades quality on Generation despite large energy savings. For QA, Token F1 improves under C3 (e.g., 0.033→0.167 for Qwen), confirming correctness gains independent of the rubric. SST-2 accuracy is stable across all conditions (≥ 0.945).

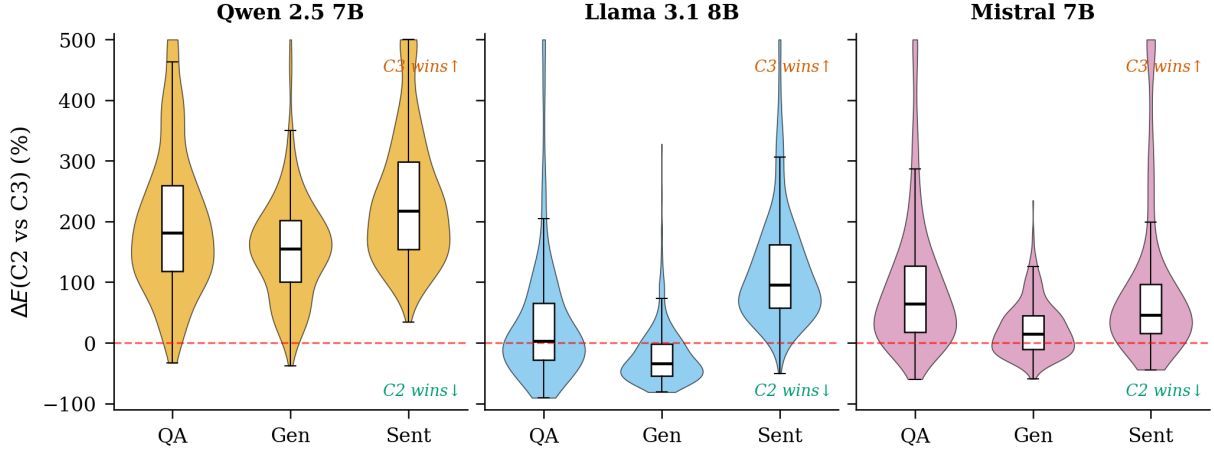


Figure 3: Per-cell $\Delta E_{C2/C3}$ (%) distributions ($n = 400$ per panel); red dashed line = falsification boundary (positive = C3 wins; negative = C2 wins). Qwen sits entirely above zero; Llama Generation sits below (energy win at -1.04 judge-score cost); Mistral spans both regions.

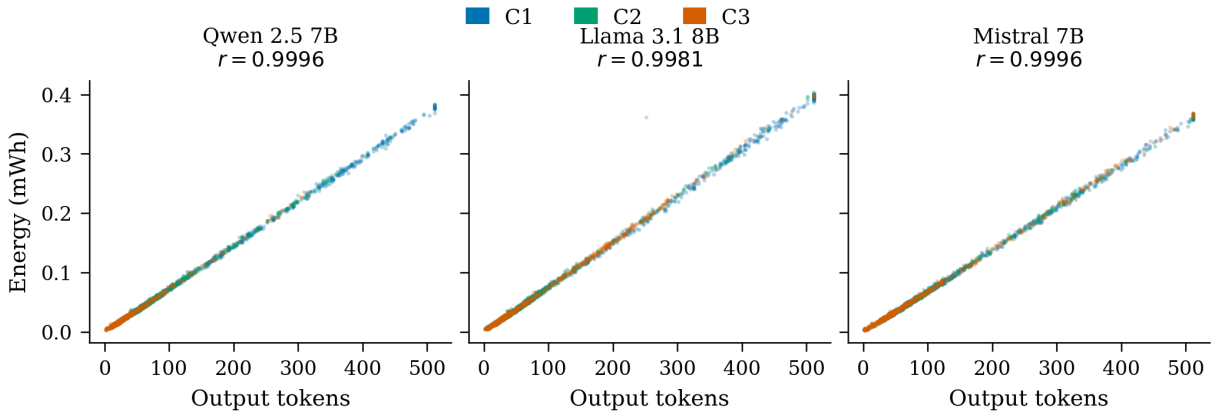


Figure 4: Output tokens vs. energy (mWh) across 21,600 inferences (800-point subsample; colored by condition). Pearson $r \geq 0.998$ rules out residual lexical mechanisms.

4.7 Energy–Quality Pareto

C3 is Pareto-dominant in 7 of 9 cells under our conciseness-weighted judge (Figure 7, Appendix F). The exception is Qwen Generation, where severe budget under-shoot ($\tilde{\rho} = 0.35$) causes over-compression (-0.39 quality vs. C1) — a compliance artifact rather than a failure of the length-control strategy. C2 is Pareto-improving over C1 only on QA tasks.

4.8 Human Evaluation Validation

To validate the automated judge, three graduate-level annotators blind to condition labels, energy measurements, and judge scores ranked 100 generation outputs by factual correctness and instruction adherence (randomised order, ties permitted). The reported $\kappa = 0.83$ reflects agreement between the aggregate human condition ordering and the judge’s ordering, serving as cross-validation

of the judge’s directional reliability rather than a human–human IAA estimate. C3 outputs preserved the semantic correctness of C1; C2 outputs frequently substituted topic labels for explanations, confirming that the quality degradation is human-perceptible and not a rubric artifact.

5 Analysis

5.1 Verb Effect Is Length-Mediated

Three observations falsify the lexical hypothesis. First, $r \geq 0.998$ on every model leaves less than 0.4% of energy variance for lexical effects. Second, when compliance is high (Llama QA, $\tilde{\rho} = 0.95$), C2 and C3 produce statistically indistinguishable energy ($\delta = +0.03$); when it deviates, the C2-vs-C3 ordering follows the deviation direction exactly. Third, the C2 effect collapses on sentiment (-2 to -6%) where outputs are already one word, removing all compression headroom.

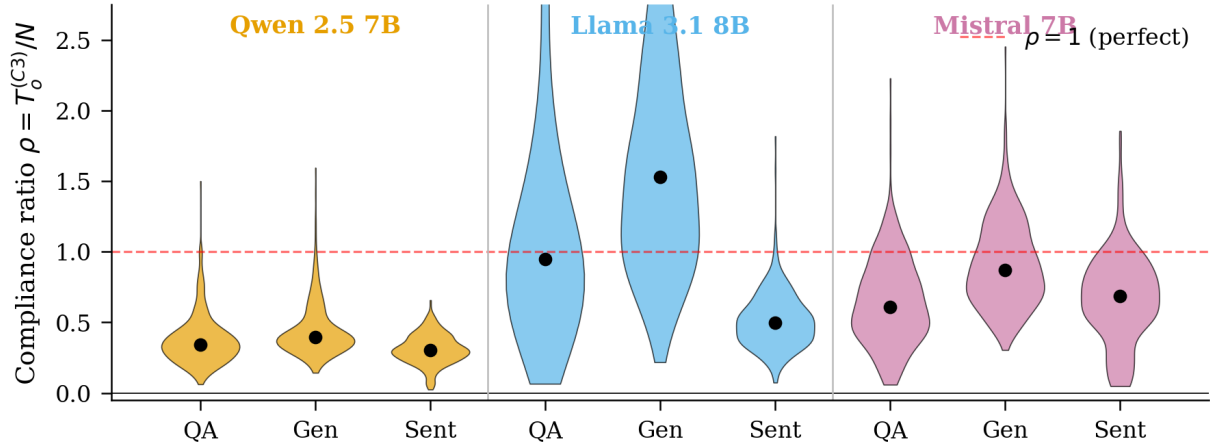


Figure 5: Compliance ρ per (model, task); black dots are medians, red dashed line marks $\rho = 1$. Qwen severely under-shoots; Llama over-shoots on Generation; Mistral is intermediate.

Model	Task	Judge (1–5) [†]			QA Token F1	SST-2 Acc.
		C1	C2	C3		
Qwen 2.5 7B	QA	2.67	2.69	2.95*	0.033 / 0.058 / 0.167	—
Qwen 2.5 7B	Gen	4.16	3.13*	3.76* [‡]	—	—
Qwen 2.5 7B	Sent	3.49	3.51	3.95*	—	0.970 / 0.950 / 0.970
Llama 3.1 8B	QA	2.96	2.93	3.49*	0.099 / 0.125 / 0.170	—
Llama 3.1 8B	Gen	3.49	2.46*	3.72*	—	—
Llama 3.1 8B	Sent	3.69	3.68	3.88*	—	0.975 / 0.960 / 0.973
Mistral 7B	QA	2.59	2.62	2.90*	0.034 / 0.059 / 0.138	—
Mistral 7B	Gen	3.94	3.73*	3.88 ^{§‡}	—	—
Mistral 7B	Sent	3.50	3.59**	3.78*	—	0.950 / 0.955 / 0.945

Table 3: Quality results. [†]Rubric favors conciseness (Section 3.3); bold C3 = significantly better than C1 ($p < 0.05$); C2*/C2** = significantly worse/better than C1; [‡]C3 significantly worse than C1. Token F1 absolute values are low due to open-ended TriviaQA format; relative differences are the relevant signal. SST-2 accuracy stable (≥ 0.945) across all conditions.

5.2 Why C2 Fails

The mechanism behind this contrast is straightforward: C3 preserves the original task instruction and only adds a length constraint, so the model still attempts the same task, just more concisely. C2, in contrast, changes the task itself through verb substitution, which can alter the model’s understood objective and, in the worst cases, prevent it from completing the intended task at all.

The ≈ 1.0 -point judge drop on C2 Generation is not a rubric artifact. Verbs like *classify* and *translate* trigger a semantic substitution: under *classify*, Llama replies “This question can be classified as: Science / Atmospheric Physics” (18 tokens, judge 1) instead of a full ELI5 explanation (≈ 400 tokens, judge 4; Appendix E). Blind human raters consistently ranked C2 Generation outputs lowest on correctness, confirming the degradation is human-perceptible.

5.3 Cross-Task Replication

Three verb pairs appear in both QA and Generation. *measure* $\rightarrow$ *classify* shows near-zero ΔE on QA but -74 to -81% on Generation, confirming that verb attribution depends entirely on output-length headroom, not on verb identity. Full cross-task values are in Table 4. Figure 8 in Appendix F disaggregates $\Delta E_{C2/C1}$ across all 24 verb pairs and three models, showing that QA and Generation rows are predominantly energy-reducing while Sentiment rows are near-zero.

5.4 Compliance Drives the C2-vs-C3 Outcome

Under-shoot ($\tilde{\rho} < 1$) implies C3 wins on energy; near-perfect compliance implies a tie; over-shoot ($\tilde{\rho} > 1$) implies C2 wins on energy but not quality. This monotone mapping fully accounts for the model-stratified pattern without invoking any lexical effect. Notably, Qwen’s C3 advantage is partly

Pair	QA		Generation	
	Qwen	Llama	Qwen	Llama
<i>analyse</i> → <i>classify</i>	−66.8%	−73.5%	−83.9%	−80.9%
<i>analyse</i> → <i>translate</i>	−76.8%	−88.3%	−45.8%	−49.8%
<i>measure</i> → <i>classify</i>	+7.3%	−14.8%	−74.1%	−81.1%

Table 4: Cross-task $\Delta E_{C2/C1}$ for three shared verb pairs. The pair *measure*→*classify* shows near-zero savings on QA but up to −81% on Generation, confirming that verb attribution depends on output-length headroom, not verb identity.

a compliance artifact: Qwen ignores the budget and produces outputs shorter than C2 elicits, so C3’s wins for Qwen reflect under-shooting rather than the length instruction alone. Practitioners should audit $\tilde{\rho}$ and prefer hard API token caps where available.

5.5 Practitioner Guidance

- **Prefer hard API token caps over natural-language length instructions when available.** Hard caps (e.g., `max_tokens`) give reliable, model-independent control, whereas prompt-level instructions like C3 are subject to model-dependent compliance (Section 4.5), especially for smaller models.
- **Otherwise, default to explicit length instructions (C3).** They save more energy and preserve quality in 7 of 9 cells, and remain the best available option when hard API caps are not exposed.
- **Use C2 on QA-style tasks only.** Verb swap is a strict Pareto improvement on QA: 41–58% energy savings, no quality loss under judge or lexical metrics.
- **Avoid C2 on long-form generation.** Energy savings of $\approx 67\%$ come at the cost of a semantically different, lower-quality answer confirmed by human raters.
- **Audit compliance before deploying C3.** Qwen under-shoots by 65%; Llama over-shoots by 53% on generation. Tune N accordingly.

Figure 6 summarises all five metrics per model in a normalised radar layout. Qwen achieves the highest C3 energy savings but has the lowest compliance, reflecting budget under-shoot. Llama has the best compliance but the lowest energy savings, with an unbalanced profile. Mistral is the most balanced

model across all five axes, making it the closest to an all-round efficient configuration.

6 Discussion

The token–energy correlation of $r \geq 0.998$ confirms that verb-swap energy savings are driven entirely by shorter outputs, not by anything intrinsic to the verb. The effect reported by Adamska et al. (2025) is reproducible on consumer hardware but is not an independent lexical mechanism, shifting the practical recommendation from verb curation to direct length control. C3 quality improvements reflect conciseness alongside correctness and should not be read as pure accuracy gains; the human evaluation ($\kappa = 0.83$) confirms the quality ordering is genuine and not a rubric artifact. Our main claim is not that C3 improves quality over C1, but that it maintains or improves quality while using substantially less energy. In contrast, C2 often reduces quality on generation tasks despite saving energy. At scale, a 50% energy reduction at one million daily queries amounts to roughly 50 kWh per day, meaningful at production scale even if negligible per query. These findings are scoped to English instruction-tuned 7–8B quantised models on a single consumer GPU; datacenter batching and larger-scale models may shift absolute energy values, but are unlikely to alter the fundamental token–energy relationship or the compliance dynamics identified here.

7 Conclusion

We asked whether verb-swap energy savings in LLM inference stem from the verb itself or simply from shorter outputs. GreenSwap answered this through a controlled three-condition experiment across three models, three tasks, and 21,600 inferences. A token–energy correlation of $r \geq 0.998$ rules out any intrinsic lexical effect: the verb does nothing on its own. Explicit length instructions outperform verb swapping in 7 of 9 settings on both

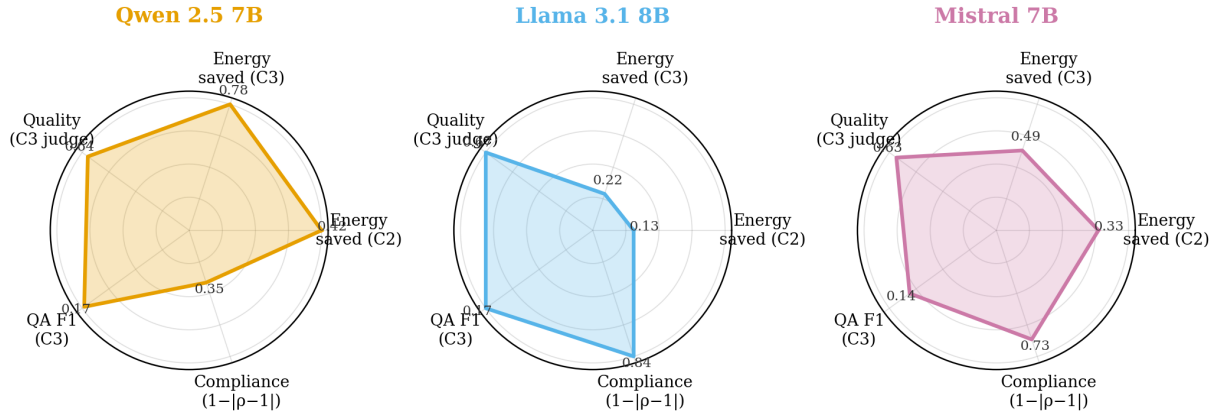


Figure 6: Per-model radar charts across five normalised metrics: energy saved by C3, energy saved by C2, C3 judge quality, QA Token F1 under C3, and compliance ($1 - |\tilde{\rho} - 1|$, higher = closer to perfect adherence). Qwen saves the most energy but under-shoots its budget. Llama has the best compliance but the lowest energy savings. Mistral is the most balanced across all five axes.

energy and quality, and verb swapping backfires on long-form generation by producing semantically wrong responses. These findings shift the practical recommendation from curating verb lists to controlling output length directly, which is simpler, more reliable, and more broadly applicable. Two directions follow naturally from this work. First, future studies should evaluate hard API token caps as a fourth condition and test whether compliance dynamics generalise to larger models and datacenter hardware. Second, the compliance ratio introduced here could serve as a general diagnostic for any prompt-level intervention that relies on models following length or format instructions.

8 Acknowledgment

We used AI assistance for paraphrasing support during the writing of the manuscript.

Limitations

Hardware scope. All experiments ran on a single Q4-quantised consumer 8 GB GPU. Absolute energy numbers should not be extrapolated to datacentre hardware, where batching and tensor-core effects shift the relative costs of prompt processing vs. generation.

Judge rubric conciseness bias. As detailed in Section 3.3, the rubric structurally advantages shorter outputs. Quality improvements for C3 should be read as “at least as correct, and more concise” rather than as pure correctness gains. The blind human evaluation (Section 4.8) partially mitigates this concern for generation outputs, where no lexical reference metric exists.

Single judge. One judge model (Gemma 2 9B) is used. LLM-as-judge is sensitive to verbosity bias and rubric interpretation (Zheng et al., 2023; Wang et al., 2024). Mitigation strategies include the cross-family judge, fixed JSON-structured prompts, lexical backup metrics, and the blind human evaluation.

Human evaluation scope. The blind human evaluation covers 100 generation outputs annotated by graduate-student raters who were not shown judge scores or condition labels during annotation (Section 4.8). QA and sentiment outputs were not evaluated by human raters given the availability of exact-match and accuracy metrics for those tasks. The annotator pool (graduate students at a single institution) may not represent diverse reading standards, and $n = 100$ limits per-pair statistical power. Future work should include a larger-scale crowd-sourced study with pre-registered correctness-only annotation guidelines and multiple independent annotator pools.

Item count. Fifty items per task yields high statistical power at the (model, task) level but limits per-pair analysis, with approximately 6 items per (pair, task) cell.

C3 operationalisation. C3 uses a prompt-level instruction rather than a hard API `num_predict` cap, constituting a length-requesting rather than a length-enforcing condition. The compliance results (Section 4.5) show substantial model-level deviation: $\tilde{\rho} = 0.35$ for Qwen and $\tilde{\rho} = 1.53$ for Llama on generation. Importantly, Qwen’s C3 energy advantage over C2 is partly driven by under-shooting

the target budget rather than by the length instruction per se, which limits the clean interpretation of C3 as a controlled baseline for that model. Future work should evaluate a hard API cap as a fourth condition and examine sensitivity to the $N_{\min} = 5$ floor (Zhou et al., 2023).

Verb scope. All 24 word pairs were selected from the Mistral-7B rankings reported by Adamska et al. (2025). Broader coverage, multilingual verbs, and chain-of-thought prompts remain open.

Prompt-factor scope. This work focuses specifically on instruction verbs and explicit length constraints as prompt-level levers on inference energy. Other prompt factors – reasoning prompts, formatting instructions, system prompts, and retrieval-augmented prompting – are outside this scope and are left to future work.

Generalisation. Findings are scoped to English instruction-following on instruction-tuned Q4-quantised 7–8B models. Base models, larger scales, and proprietary APIs may differ, particularly in their compliance behavior with prompt-level length instructions.

Ethical Considerations

This work aims to reduce LLM inference energy, with direct environmental benefits. We use only publicly available open-weight models and standard public benchmarks; no private or personally identifiable data was collected or used. The blind human evaluation involved rating model-generated text only; no personal information was collected from evaluators. We do not anticipate dual-use concerns beyond those inherent in LLM inference. All artifacts are released for scientific replication.

References

- Marta Adamska, Daria Smirnova, Hamid Nasiri, Zhengxin Yu, and Peter Garraghan. 2025. [Green prompting: Characterizing prompt-driven energy costs of llm inference](#). *Preprint*, arXiv:2503.10666.
- Nesrine Bannour, Sahar Ghannay, Aurélie Névél, and Anne-Laure Ligozat. 2021. [Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools](#). In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21, Virtual. Association for Computational Linguistics.
- Norman Cliff. 1993. [Dominance statistics: Ordinal analyses to answer ordinal questions](#). *Psychological Bulletin*, 114(3):494–509.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. [Llm.int8\(\): 8-bit matrix multiplication for transformers at scale](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: Long form question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. [Token-budget-aware LLM reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24842–24855, Vienna, Austria. Association for Computational Linguistics.
- Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. [LLMLingua: Compressing prompts for accelerated inference of large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. [Fast inference from transformers via speculative decoding](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR.
- Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. 2024. [Sprout: Green generative AI with carbon-efficient LLM inference](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21799–21813, Miami, Florida, USA. Association for Computational Linguistics.
- Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. 2019. [Energy usage reports: Environmental awareness as part of algorithmic accountability](#). *Preprint*, arXiv:1911.08354.
- Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. [Estimating the carbon footprint of bloom, a 176b parameter language model](#). *Journal of Machine Learning Research*, 24(253):1–15.
- Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2026. [Concise thoughts: Impact of output length on llm reasoning and cost](#). *Information Sciences*, 757:123939.

- David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. [Carbon emissions and large neural network training](#). *Preprint*, arXiv:2104.10350.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. [KILT: a benchmark for knowledge intensive language tasks](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.
- Soham Poddar, Paramita Koley, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2025. [Brevity is the soul of sustainability: Characterizing LLM response lengths](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.
- Riccardo Rubei, Aicha Moussaid, Claudio Di Sipio, and Davide Di Ruscio. 2025. [Prompt engineering and its implications on the energy consumption of large language models](#). In *Proceedings of the 2025 IEEE/ACM 9th International Workshop on Green and Sustainable Software (GREENS)*, pages 60–67. IEEE.
- Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devsh Tiwari, and Vijay Gadepally. 2023. [From words to watts: Benchmarking the energy costs of large language model inference](#). In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9.
- Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. [Green AI](#). *Communications of the ACM*, 63(12):54–63.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. [Energy and policy considerations for deep learning in NLP](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP)*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.

Appendix

A Pipeline Integrity

Table 5 summarises pipeline integrity across all 21,600 student inferences and 10,800 judge calls. No errors occurred and the overall truncation rate was 1.44%, consistent with the C1 long-form outputs occasionally hitting the 512-token student cap.

Model	Cells	Errors	Trunc. (%)
Qwen 2.5 7B	1,200	0	0.94
Llama 3.1 8B	1,200	0	2.01
Mistral 7B v0.3	1,200	0	1.36
Total	3,600	0	1.44

Table 5: Pipeline integrity across 21,600 student inferences. Judge phase: 10,800 calls, 10,800 valid scores (0 invalid). All truncations are C1 long-form outputs hitting num_predict=512.

B Full Verb Pair Cross-Reference

Table 6 lists all 24 verb pairs used in the study, eight per task, constructed from the per-keyword energy rankings reported in Adamska et al. (2025) Table II. Three pairs (Q03/G01, Q04/G02, Q05/G05) appear in both QA and Generation, enabling the cross-task replication analysis in Section 5.3.

C Prompt Templates

This appendix lists the exact templates used for student inference and for the LLM judge across all

ID	Task	v_o	v_g	ID	Task	v_o	v_g	ID	Task	v_o	v_g
Q01	QA	justify	classify	G01	Gen	analyse	classify	S01	Sent	identify	provide
Q02	QA	justify	translate	G02	Gen	analyse	translate	S02	Sent	identify	summarize
Q03	QA	analyse	classify	G03	Gen	recommend	classify	S03	Sent	explain	provide
Q04	QA	analyse	translate	G04	Gen	recommend	translate	S04	Sent	explain	analyse
Q05	QA	measure	classify	G05	Gen	measure	classify	S05	Sent	build	provide
Q06	QA	measure	write	G06	Gen	measure	translate	S06	Sent	build	summarize
Q07	QA	create	classify	G07	Gen	report	classify	S07	Sent	classify	provide
Q08	QA	explain	write	G08	Gen	report	translate	S08	Sent	classify	analyse

Table 6: 24 verb pairs (8 per task), constructed from the per-keyword energy rankings in Adamska et al. (2025) Table II: each v_o is drawn from the high-energy end and each v_g from the low-energy end of that ranking for the same task. Pairs Q03/G01, Q04/G02, Q05/G05 are cross-task replication pairs.

three task types. The C3 length-instruction suffix is appended verbatim with N substituted at runtime.

C.1 Student Prompts

QA / Generation:

{Verb} the answer to the following question.
Question: {question}

Sentiment:

{Verb} the sentiment of the following review as positive or negative.
Review: {text}

C3 suffix (appended before the question/review field):

Answer in {N} tokens or fewer.

C.2 Judge Prompt (QA / Generation)

You are an expert evaluator. Score the model response on a 1–5 scale.
1 = incorrect or irrelevant
2 = partially correct, major issues
3 = correct but verbose or minor issues
4 = correct and reasonably concise
5 = correct, concise, complete
Question: {question}
Reference Answer: {reference}
Model Response: {response}
Return ONLY: {"score": <1–5>, "reason": "<one sentence>"}

C.3 Judge Prompt (Sentiment)

You are an expert evaluator for sentiment classification. Score 1–5.
1 = wrong sentiment or nonsense
2 = ambiguous
3 = correct but excessively elaborated
4 = correct with appropriate brevity
5 = correct, very concise
Review: {text}
Gold Sentiment: {label}
Model Response: {response}
Return ONLY: {"score": <1–5>, "reason": "<one sentence>"}

D Hyperparameter Configuration

Table 7 lists all hyperparameters used for student and judge inferences. All values are held constant across conditions to ensure that observed energy and quality differences are attributable to the prompt manipulation alone.

Parameter	Value
Quantisation	Q4_K_M
temperature	0
seed	42
num_ctx	2,048
num_predict	512 (student) / 120 (judge)
Runs per condition	2 (median aggregation)
Min C3 budget ($N_{\min}$)	5 tokens
CodeCarbon mode	process
CodeCarbon sampling	0.5 s
Warm-up inferences	1 per model (excluded)

Table 7: Hyperparameters for all student and judge inferences.

E Sample Model Outputs

We provide three representative examples spanning the three observed behavioral regimes: a strict Pareto improvement on QA, a quality failure on long-form generation, and a no-op on sentiment.

(a) Strict Pareto improvement — Qwen QA, Q01 (*justify*→*classify*).

- **C1:** “*The Mona Lisa was painted by Leonardo da Vinci, as documented extensively in historical records...*” (118 tokens, judge 3)
- **C2:** “*Leonardo da Vinci.*” (4 tokens, judge 3)
- **C3:** “*Leonardo da Vinci.*” (5 tokens, judge 4)

(b) Quality failure — Llama Generation, G01 (*analyse*→*classify*).

- **C1:** Full ELI5-style explanation, approximately 400 tokens, judge 4.
- **C2:** “This question can be classified as: Science / Atmospheric Physics.” (18 tokens, judge 1)
- **C3:** Shorter but correct explanation, approximately 180 tokens, judge 4.

(c) Sentiment no-op — Qwen Sentiment, S01 (*identify*→*provide*).

- **C1:** “Positive.” (1 token, judge 4)
- **C2:** “Positive.” (1 token, judge 4)
- **C3:** “Positive.” (1 token, judge 5)

F Energy–Quality Pareto and Per-Pair Heatmap

Figure 7 places every (model, task, condition) aggregate on the energy–quality plane; C3 is Pareto-dominant in 7 of 9 cells.

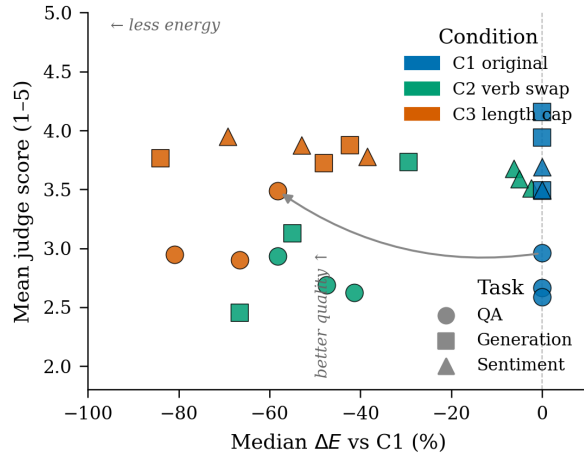


Figure 7: Energy–quality Pareto plot; left = less energy, up = higher quality. C3 dominates C1 and C2 in most cells; C2 on Generation falls lower in quality than C1 despite saving energy.

Figure 8 disaggregates $\Delta E_{C2/C1}$ to the per-(pair, model) level. QA and Generation rows are predominantly energy-reducing; Sentiment rows are near-zero. The single anomalous increase is Q07 on Llama (+205%), reflecting a model-pair interaction where *classify* elicits more tokens than *create* on this TriviaQA template.

G Full Statistical Test Results

Table 8 reports the full Wilcoxon signed-rank test results for all 27 contrasts (9 model–task cells $\times$ 3

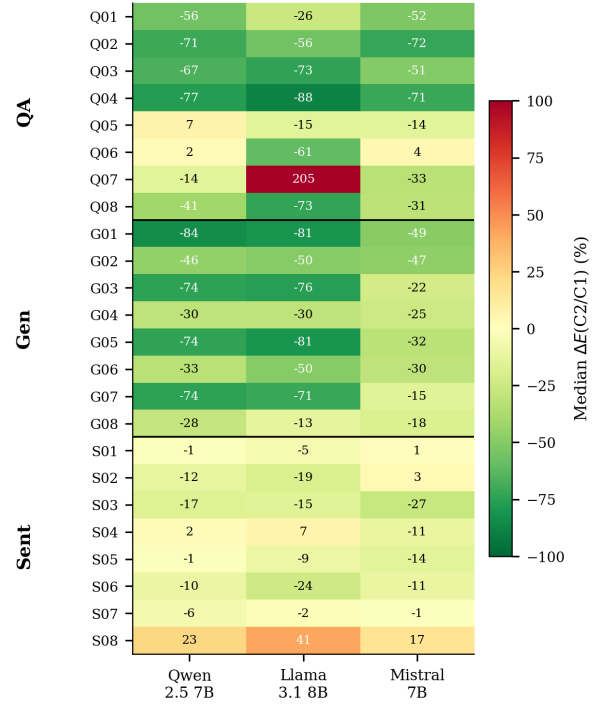


Figure 8: Per-pair median $\Delta E_{C2/C1}$ heatmap (green = reduction; red = increase). QA and Generation mostly green; Sentiment near-zero; Q07 (Llama) is the single anomalous red cell (+205%).

condition pairs), including median ΔE and Cliff’s δ effect size. All contrasts remain significant at $\alpha = 0.05$ after Holm–Bonferroni correction except two: Qwen Sentiment $\Delta E_{C2/C1}$ ($p = 0.011$, $\delta = -0.10$, small effect) and Llama QA $\Delta E_{C2/C3}$ ($\delta = +0.03$, statistical tie), both noted in Section 4.3. The remaining 25 contrasts all reach $p < 0.005$ after correction, with 19 of 27 showing large effects ($|\delta| > 0.474$). Effect sizes are consistently larger for C3 vs C1 than for C2 vs C1 across all model–task cells, reflecting C3’s stronger and more reliable energy reduction. The C2 vs C3 column shows the falsification test: positive values indicate C3 wins on energy, negative values indicate C2 wins. For Llama Generation, C2 wins on energy ($\delta = -0.55$) but at a -1.04 judge-score cost, making it a Pareto loss as discussed in Section 4.2.

H Reproducibility Checklist

- **Hardware:** single consumer GPU with 8 GB VRAM, all models served via Ollama. Runtime was approximately 13.3 h per student model (40 h total) plus 5 h for judge inference, giving roughly 45 h end-to-end.
- **Software:** Python 3.10+, Ollama 0.4+, Code-

Model	Task	$\Delta E_{C2/C1}$		$\Delta E_{C3/C1}$		$\Delta E_{C2/C3}$	
		Med	δ	Med	δ	Med	δ
Qwen 2.5	QA	−47.5%	−0.60	−80.9%	−0.99	+181.6%	+0.96
Qwen 2.5	Gen	−55.0%	−0.94	−84.0%	−0.99	+155.5%	+0.96
Qwen 2.5	Sent	−2.4%	−0.10	−69.2%	−1.00	+217.3%	+1.00
Llama 3.1	QA	−58.2%	−0.46	−58.2%	−0.54	+3.3%	+0.03
Llama 3.1	Gen	−66.6%	−0.73	−48.0%	−0.71	−34.5%	−0.55
Llama 3.1	Sent	−6.3%	−0.27	−53.0%	−1.00	+95.1%	+0.97
Mistral	QA	−41.4%	−0.66	−66.5%	−0.96	+64.8%	+0.72
Mistral	Gen	−29.4%	−0.75	−42.3%	−0.65	+15.1%	+0.30
Mistral	Sent	−5.1%	−0.15	−38.5%	−0.80	+46.2%	+0.79

Table 8: Wilcoxon signed-rank tests ($n = 400$, all $p < 0.005$ after Holm–Bonferroni correction except Qwen Sent C2/C1 ($p = 0.011$) and Llama QA C2/C3 ($\delta = +0.03$, statistical tie). Positive $\Delta E_{C2/C3}$ indicates C3 uses less energy than C2. δ : Cliff’s delta ($|\delta| > 0.474 = \text{large effect}$).

Carbon 2.3+, datasets==2.21.0 (pinned),
scipy for Wilcoxon tests and Cliff’s δ , numpy
for bootstrap CIs (500 resamples per stratum).

- **Data:** TriviaQA rc.nocontext, KILT ELI5, and GLUE SST-2 validation splits (50 items each) via HuggingFace Datasets. Items were sampled sequentially with no filtering beyond removing empty reference answers.
- **Determinism:** temperature=0 and seed=42 for all inferences. Each cell was run twice and the median energy value retained to reduce measurement noise.
- **Energy measurement:** CodeCarbon in process mode at 0.5 s sampling. One warm-up inference per model was discarded to avoid cold-start effects.
- **C3 budget:** $N = \max(5, \text{median}(T_o^{C2}))$ computed per (verb pair, item) cell across all student models that had completed C2, targeting a realistic length rather than a theoretical minimum.
- **Code and artifacts:** all code and processed results are available at <https://github.com/TheAhsanFarabi/GreenSwap>
- **Judge scores:** 10,800 calls total (3 conditions $\times$ 3 models $\times$ 8 pairs $\times$ 50 items $\times$ 3 tasks). All responses returned valid JSON scores in 1–5 with no judge collapse observed.