

What a Word Limit Removes: Length Constraints Look Free, Costly or Helpful Depending on the Span They Bind, the Scorer That Reads the Answer and the Items Chosen, and Why Compliance Looks Orthogonal to Accuracy in Only One Kind of Design

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Published work on telling a language model to be brief disagrees about almost everything. One benchmark reports that obeying a word limit and answering accurately are statistically unrelated, and that alignment for helpfulness is what makes models overrun the limit. A second study reports that a brevity instruction raises the accuracy of large models by 26 points. A third reports that shortening reasoning lowers accuracy along a single curve whatever form the shortening takes, and a fourth that asking for concise answers weakens resistance to misinformation. In this analysis the author reviews and synthesises published results to show that these findings are compatible once three things are stated that the studies usually leave implicit: which span of text the limit binds (the prompt, the reasoning, or the answer), which scoring rule reads the shortened output, and which items the comparison was run on. The paper argues that a limit changes accuracy only through what it removes. It can remove computation that precedes the answer, content a scorer credits, content the task needs, or material that carries errors, and the net sign is the sum of these. Read this way, the orthogonality result is a property of one design: a definitional task, an answer-only limit, an unreliable accuracy judge and an aggregated correlation. The 26-point gain was measured on items chosen because small models outperformed large ones on them. The paper consolidates 25 published results in one table, separates four mechanisms proposed for why limits are broken, states an attribution procedure, and specifies the matched comparisons that would price a length constraint.

Keywords: length constraints; instruction following; conciseness; chain-of-thought compression; verbosity bias; LLM-as-a-judge; evaluation methodology; measurement validity

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Length is the most common constraint a user puts on a language model and one of the least understood. Instruction-following benchmarks treat a word limit as a verifiable instruction alongside keyword and format requirements (Zhou et al., 2023; Jiang et al., 2023b; Yao et al., 2023). Deployment guides treat brevity as a cost lever, since output tokens are billed and slow. Alignment research treats length as a confound, because preference-trained models and the judges that score them both drift toward longer text (Singhal et al., 2023; Dubois et al., 2024). Three communities therefore study the same instruction and ask different questions about it. Can models follow it? What does following it cost? Does it distort measurement?

A December 2025 benchmark pulls the three questions together. Baxi (2025) introduces the Compression-Decay Comprehension Test, which holds a 35-word answer limit fixed while the prompt explaining a scientific concept is rewritten at five lengths, from about 2 words to about 135. Two scores are given: constraint compliance and semantic accuracy. Its headline claims are that compliance traces a U-shape across prompt lengths, worst near 27 words; that compliance and

accuracy are statistically orthogonal; and that the helpfulness behaviours instilled by alignment are the main cause of limit violations. In its supporting experiment, removing helpfulness language from the system prompt is reported to raise compliance by 598 percent on average.

The surrounding literature does not sit easily with any of these claims. Hakim (2026) reports that a brevity instruction raises large models' accuracy by 26.3 points on a set of benchmark items and reverses their deficit against small models on two benchmarks. Lee et al. (2025) report that across 31 prompts that shorten chain-of-thought, accuracy falls along a near-universal curve set by length alone. Le Jeune et al. (2025) report that system prompts asking for concise answers lower resistance to misinformation in 11 of 17 models. Zhang, Das and Zhang (2024) report that when models are told to answer concisely and do not, the long answers are less accurate than the short ones, a direct association between non-compliance and error. Brevity is therefore reported as free, as costly and as beneficial. Compliance is reported as orthogonal to accuracy by one study and as associated with it by another.

This paper argues that these results do not contradict one another. They measure different things under one name. Its central claim is simple to state: a length limit

changes accuracy only through what it removes. What it removes depends on three design choices that published studies rarely report together:

First, the span the limit binds. A limit can bind the prompt (input compression), the reasoning the model writes before its answer, or the answer itself. These are different interventions with different expected effects.

Second, the scorer that reads the output. Exact-match extraction, recall against a reference, atomic-fact precision, pairwise LLM judges and rubric judges each respond to length in their own way. Some reward it, some penalise it, and one family of designs makes a violated limit an automatic loss.

Third, the items and models compared. A brevity effect measured on items chosen because a model failed them is not an estimate of the average effect.

Once these three are fixed, the apparent conflict separates into findings about different quantities. The paper then takes up the second dispute, why models break length limits at all. It separates four mechanisms that the literature proposes and shows that the evidence assigns the anchor's helpfulness explanation a narrower role than claimed.

The genre is analysis. In this analysis the author reviews and synthesises published results to show where the disagreement lives and what would settle it. Every number below is credited to the study that reported it, and where the paper does arithmetic on a source's reported values, it says so. The paper's own contributions are the span-scorer-selection decomposition, the four-channel account of what a limit removes, the consolidated evidence in Table 1 and Table 2, the attribution procedure in Algorithm 1, and the matched design in Section 12.

The most obvious objection should be answered at once. A reader may say that the anchor is a single-author preprint with internal inconsistencies, and that a paper built around it is over-investing in a weak source. The answer is that the anchor matters as a clean statement of a position that many deployment discussions take for granted: being brief is a formatting matter separate from being right, and the main thing preventing brevity is a trained wish to be helpful. Both halves of that position have stronger and weaker published support, and the paper's interest is in the position, not the preprint. A second objection is that this paper might restate an earlier analysis of whether longer chain-of-thought makes reasoning more reliable. It does not. That question concerns how much deliberation helps. This one concerns what an externally imposed length limit does to accuracy and to measurement, and most of its evidence concerns answer-length and scoring rather than reasoning length.

The paper proceeds as follows. Section 2 sets out what the anchor's design fixes. Section 3 defines the three

spans and three enforcement modes. Sections 4 and 5 assemble the evidence that brevity costs accuracy and the evidence that it is free or helps. Section 6 states the four removal channels that reconcile them. Section 7 shows how the scoring rule decides whether compliance and accuracy look related. Section 8 takes up selection. Section 9 separates the four explanations of limit violations. Section 10 consolidates the evidence, Section 11 states the attribution procedure, and Section 12 specifies the comparisons that would settle the open part. Sections 13 to 15 state the case against the paper, what is not known, and its limitations.

2. The anchor study and what its design holds fixed

The anchor deserves a precise description, because its conclusions are claims about its design and no wider than it.

Task and constraint. Models are asked to explain one of eight concepts: *modus ponens*, recursion, derivative, photosynthesis, natural selection, the harm principle, impressionism and theory of mind. The answer must be exactly 35 words, and that target is held fixed across conditions (Baxi, 2025, Section 3.1). The manipulated variable is the length of the prompt. A detailed prompt of about 135 words, with context, examples and formatting instructions, is rewritten by GPT-5.1 into versions of about 69, 27, 13 and 2 words. The rewriting instruction asks the compressor to keep the 35-word requirement while cutting context and examples (Section 3.3). Nine models are evaluated at temperature zero.

Scoring. Both outcomes are scored by a jury of three LLM judges on 0-10 rubrics, and the scores are averaged. On the compliance rubric, 10 means exactly 35 words, 8-9 means within one to three words, 6-7 within four to seven, 4-5 within eight to fifteen, and 0-3 a larger deviation. Semantic accuracy runs from "complete, accurate, and well-structured" at 10 to "mostly incorrect or irrelevant" at the bottom. On ratings discretised at a 0.7 threshold, the source reports Fleiss' kappa of 0.90 for compliance, 0.25 for accuracy and 0.19 for a third dimension, functional completeness (Section 3.7).

Results. Mean compliance is 8.12 at the shortest prompt, 6.54 at the 27-word prompt and 8.03 at the longest. The U-shape appears in 70 of 72 model-concept combinations. Across what the paper calls 81 experiments, the correlation between compliance and accuracy is $r = 0.193$ with a 95 percent interval of -0.025 to 0.396 and $p = 0.084$. Accuracy rises slowly with prompt length, by 0.090 on average from the shortest to the longest prompt. Three reasoning models reach mean compliance of 8.20 against 7.30 for the other six.

The ablation. The supporting experiment re-runs the 27-word condition with system prompts from which phrases encouraging comprehensive, detailed and helpful

responses have been removed, while the 35-word specification stays (Section 4.6.1). Compliance on a 0-1 scale rises from per-model baselines of 0.08 to 0.29 to ablated values of 0.95 to 1.00. Improvement appears in 71 of 72 trials, and 57 of 72 reach perfect compliance. The source's own prediction for this experiment had been an improvement of 40 to 50 percent (Section 5.2).

Six features of this design limit what it can show, and each matters later.

First, **the limit binds only the answer, on a task that needs no reasoning before the answer.** Explaining photosynthesis in 35 words requires recall and exposition, not multi-step derivation. Section 6 argues that this is precisely the condition under which a limit removes the least accuracy-bearing computation. An orthogonality result here says little about limits on tasks where the words are the work.

Second, **"compression" means something other than its usual sense.** In the prompt-compression literature the term refers to cutting inputs of thousands of tokens, such as retrieved documents and demonstrations, with methods like LLMLingua, LongLLMLingua, LLMLingua-2, Selective Context and learned summary vectors (Jiang et al., 2023a; Jiang et al., 2024; Pan et al., 2024; Li et al., 2023; Chevalier et al., 2023). In that regime compression can raise accuracy: LongLLMLingua reports a gain of up to 21.4 percent on NaturalQuestions with about four times fewer tokens. A comparative study reports that extractive compression often outperforms other methods and allows up to tenfold reduction with little accuracy loss (Jha et al., 2024). The anchor's prompts run from 2 to 135 words. Varying them changes how fully the instruction is stated, not how much evidence the model reads. Its findings are about instruction wording, and they do not transfer to input compression in either direction.

Third, **compliance was judged rather than counted.** Word count is the one outcome in the study that a few lines of code can compute exactly, and Yuan et al. (2024) do compute it, with a fixed tokenizer. The anchor instead asks three LLM judges to place each answer in a rubric band. Models are poor at counting words in their own outputs (Zhang et al., 2025; Xiao et al., 2026), and transformer counting degrades with length and depends on tokenisation (Zhang, Cao and You, 2024). Agreement among judges therefore shows that the judges agree. It does not show that they counted correctly. Section 7 returns to why a kappa of 0.90 on thresholded ratings cannot carry the validity claim the source places on it.

Fourth, **the accuracy measure is unreliable by the source's own statistic.** A kappa of 0.25 is "fair" on the conventional banding (Landis and Koch, 1977). An unreliable measure lowers any correlation computed with it, so a weak correlation between compliance and accuracy is the expected result whether or not the underlying quantities are related (Spearman, 1904).

Section 7 works this through.

Fifth, **the shortest condition cannot state the full instruction.** A prompt of two or three words can at most name the concept and a number, and the source does not show the $c=0.0$ prompts it used. The source's example of a maximally compressed prompt is " $F=ma$ ". The source explains high compliance at this level as models falling back on a naturally terse output mode that "happens to satisfy length requirements" (Section 5.1). On that account, the left arm of the U measures how close a model's default answer length lies to 35 words. It does not measure compliance with a stated limit. The U-shape may still be real for the other four levels, but the claim that models comply best when the prompt is shortest compares two different things.

Sixth, **the ablation removed a competing instruction, not a trained behaviour.** No weights were changed. The baseline system prompts asked for comprehensive and detailed answers and also for exactly 35 words. The ablated prompts kept only the second request. A large rise in compliance is what one would expect when a contradictory instruction is withdrawn. It is evidence that models resolve conflicting length instructions toward elaboration under that prompt. It is not evidence about what preference training did to the model, which Section 9 takes up. The ablation table also reports values that do not match the main study. Its baselines at the 27-word level lie between 0.08 and 0.29 on a 0-1 scale, while the main study reports a mean of 6.54 out of 10 at the same level. Its nine models include six (gpt-oss-120b, mistral-medium-2505, claude-haiku-4-5, phi-4, Llama-4-Maverick and grok-4-fast-non-reasoning) that are absent from the main study's list, which shares only o3, o4-mini and GPT-5 with it. Even the ablation's own ranges differ within the source: the per-model table gives baselines of 0.08 to 0.29 and ablated values of 0.95 to 1.00, while the figure caption reports "from 0.06–0.26 (baseline, red) to 0.96–1.00 (ablated, green)" (Section 4.6.2). The paper does not reconcile the two experiments. They are reported here as stated and treated as separate studies.

None of this means the anchor's observations are wrong. A judge-scored compliance trough at medium prompt length, a weak correlation, and a large compliance gain when a contradictory instruction is withdrawn are all plausible findings. What they do not support is the general reading the source gives them: that brevity and accuracy are independent dimensions, and that alignment-trained helpfulness is the main cause of limit violations.

3. Three spans and three enforcement modes

Two distinctions organise the rest of the paper. They are stipulations, and they are stated flatly.

Spans. A length constraint binds one of three spans.

S1, the input span. The prompt or context is shortened before the model reads it. Prompt compression methods act here (Jiang et al., 2023a; Pan et al., 2024), and so does the anchor's manipulation in a much smaller form.

S2, the reasoning span. The text the model writes before committing to an answer is shortened. That text may be visible chain-of-thought or a hidden reasoning trace. Concise chain-of-thought, Chain of Draft, token-budget prompting and trained reasoning-length control act here (Renze and Guven, 2024; Xu et al., 2025a; Han et al., 2024; Aggarwal and Welleck, 2025).

S3, the answer span. The text the user reads as the answer is shortened, with no change to any reasoning before it. The anchor's 35-word explanations, Phare's concise system prompts, the length-instructed AlpacaEval items, and summary-length control act here (Baxi, 2025; Le Jeune et al., 2025; Yuan et al., 2024; Kikuchi et al., 2016).

For a non-reasoning model answering without chain-of-thought, S2 and S3 are the same text. That merger is the source of much confusion. A 50-word limit on a GSM8K solution, as in Hakim (2026), binds both the derivation and the answer. A 35-word limit on an explanation binds only the answer. For a reasoning model with a hidden trace, an S3 limit leaves S2 untouched, and the model can draft, count and revise inside the trace. LIFE Bench observes this behaviour in a case study of a thinking model with a visible trace, which counts words in its reasoning and regenerates drafts that miss the target, and finds the calibration only partly effective (Zhang et al., 2025, Section 5.3). This alone could account for the anchor's finding that reasoning models comply better, without any appeal to differences in preference training, and the anchor's design does not test between the two. Section 9 returns to it.

Enforcement. A constraint can be enforced in three ways.

E1, soft instruction. The limit is stated in the prompt and the model may ignore it. Most of the literature works here.

E2, hard truncation. Generation is cut at a token budget, as in budget forcing and output-budget studies (Muennighoff et al., 2025; Sun et al., 2025).

E3, trained control. The model is trained to respect stated lengths, through length-aware preference data, length tokens, length-difference position encodings, or reinforcement learning with length rewards (Yuan et al., 2024; Li et al., 2024; Song et al., 2024; Butcher et al., 2024; Jie et al., 2023).

Under E1 there are two outcomes to measure: whether the limit was met, and whether the answer was right. Under E2 the limit is always met, so only accuracy varies, and it varies through what the cut removed. Under E3 compliance becomes a training target, and the question is whether accuracy was traded for it. A finding

about one enforcement mode does not transfer to another. Sun et al. (2025) report that early stopping beats cutting generation at the budget on every dataset and prompt style they test. They also report that under strict budgets the ranking of model sizes need not be monotone, so a smaller model can beat a larger one. Hakim (2026), under soft instructions on a selected item set, reports the opposite reversal: brevity lets the larger models win. Both reversals can hold, because they concern different enforcement modes and different items.

4. The evidence that brevity costs accuracy

The cost side is strongest where the limit binds reasoning that the task needs.

A length-accuracy curve for reasoning. Lee, Che and Peng (2025) prompt models with 31 instructions that shorten chain-of-thought. Some limit words, characters, tokens or steps. Others ask for bullet points, numerals only, no proper grammar, or reasoning in Chinese. On MMLU-Pro Math, Table 2 of their paper reports that GPT-4o moves from 81 percent at 586 mean tokens with default chain-of-thought to 80 percent at 415 tokens with "be concise", 79 percent at 222 tokens with numerals only, and 42 percent at 10 tokens with no chain-of-thought. Claude 3.5 Sonnet moves from 81 percent at 324 tokens to 78 percent at 227 and 48 percent at 8 tokens. The authors' central observation is that nearly all 31 prompts fall on one trade-off curve between length and accuracy. They explain this with a per-question threshold they call token complexity: below some number of tokens a question is not solved, whatever the form of the reasoning. Their threshold classifier predicts benchmark accuracy within 6 percent error. Figure 1(b) re-plots their Table 2. Their finding is about S2 under E1, and it says that the reasoning span is computation: remove enough of it and accuracy falls.

Chain of Draft's own tables. Xu et al. (2025a) propose limiting each reasoning step to about five words. They report that the method matches or beats chain-of-thought accuracy on some tasks while using as little as 7.6 percent of the tokens. Their GSM8K tables show a cost that the headline does not. With few-shot examples, GPT-4o falls from 95.4 to 91.1 percent and Claude 3.5 Sonnet from 95.8 to 91.4 percent (Table 1). Without examples, GPT-4o falls from 94.8 to 84.4 percent and Claude 3.5 Sonnet from 90.4 to 65.5 percent (Table 5). For small models the falls are larger than the few-shot falls for the frontier models: Zoom-SLM-2.3B goes from 77.7 to 50.9 percent, Llama-3.2-3B from 70.7 to 52.5 percent and Qwen2.5-3B from 59.1 to 43.1 percent (Table 6). The authors attribute the zero-shot and small-model losses to the absence of draft-style reasoning in training data. The sports-understanding task shows the opposite pattern, with a 92.4 percent token reduction for Claude 3.5 Sonnet and no accuracy loss reported there. Within one paper, then, the same instruction is free on a

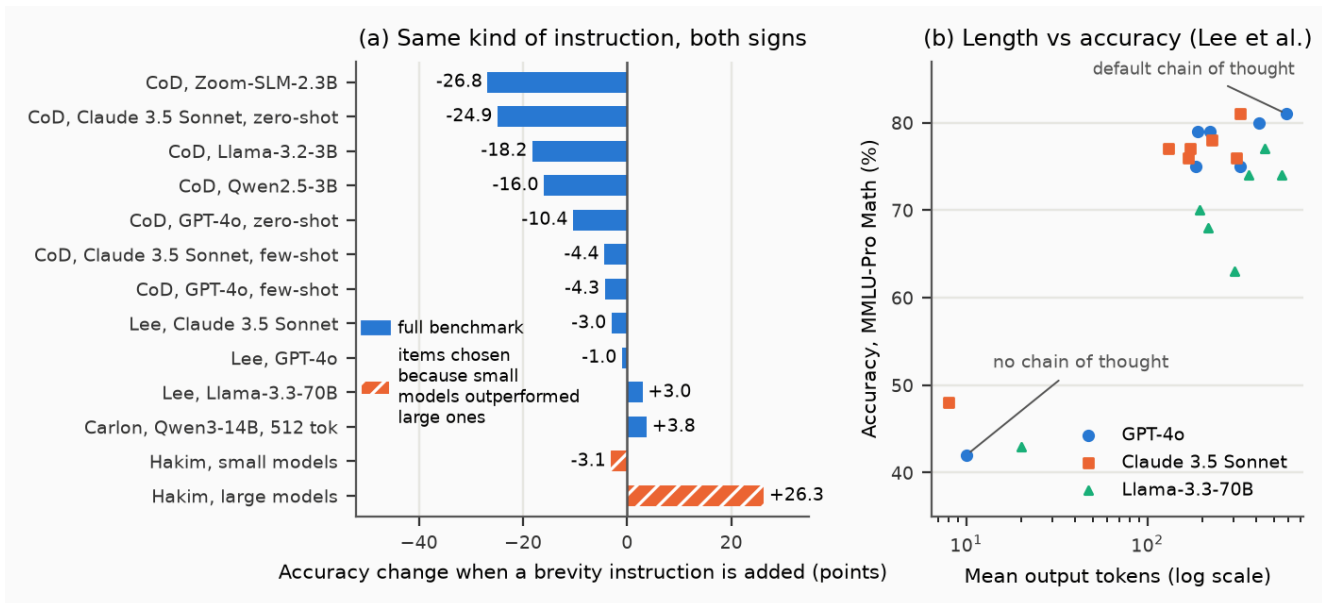


Figure 1. (a) Change in accuracy, in percentage points, when a brevity instruction is added, as reported or as computed by the author from each source's reported accuracies: Chain of Draft against chain-of-thought on GSM8K (Xu et al., 2025a, Tables 1, 5 and 6); "be concise" against default chain-of-thought on MMLU-Pro Math (Lee et al., 2025, Table 2); a concise answer-early instruction at a 512-token horizon (Carlson et al., 2026); and Brief against Control on items drawn from the 115 on which large models underperformed small ones, the source's paired test covering 96 of them (Hakim, 2026, Section 4.3; hatched). (b) Accuracy against mean output tokens for seven chain-of-thought prompts and three models on MMLU-Pro Math (Lee et al., 2025, Table 2). The same kind of instruction moves accuracy in both directions; the largest gain is on selected items.

commonsense task and costs up to about 27 points on arithmetic (Zoom-SLM-2.3B, Table 6).

Concise chain-of-thought. Renze and Guven (2024) add "be concise" to a chain-of-thought prompt on a multiple-choice benchmark. Response length falls by 48.70 percent. Across all domains the change in accuracy is not significant for GPT-3.5 or GPT-4. For GPT-3.5 on the two math exams, however, they report a performance penalty of 27.69 percent, significant by a Mann-Whitney test, while GPT-4 shows no significant math penalty. They note that GPT-4's median accuracy sits at the ceiling, which may hide an effect. Nayab et al. (2024) propose a constrained variant with explicit length targets and report it effective across models on their own correct-conciseness metrics.

Format constraints that reorder computation. Tam et al. (2024) study a different constraint with a revealing mechanism. Under JSON mode, 100 percent of GPT-3.5-Turbo's responses on a letter-concatenation task put the answer field before the reasoning field. Chain-of-thought thus became direct answering, and accuracy fell. Across reasoning tasks, stricter format restrictions produced larger losses. On classification tasks, JSON mode was competitive or better, which the authors attribute to a narrowed answer space. The constraint itself was not about length, but the mechanism is the one that matters here: a constraint costs accuracy when it removes or reorders the computation before the answer.

Hard budgets. Under strict output budgets, Sun et al. (2025) test 30 models and report that the best model size and prompt style change with the budget. Aggarwal and

Welleck (2025) train a reasoning model, L1, to respect a stated length, and present the result as a smooth trade between compute and accuracy. Muennighoff et al. (2025) use the converse lever: forcing a model to keep thinking by appending "Wait" improves accuracy on competition math, which only makes sense if the reasoning span carries accuracy. Elastic Reasoning gives separate budgets to thinking and solution and trains the model to cope with truncated thinking (Xu et al., 2025b). That design treats the reasoning span as something that can be cut and the answer span as something to protect. The wider literature on scaling test-time compute takes the same view of reasoning tokens as a resource (Snell et al., 2024).

Answer-span limits that remove what the task needs.

The most direct evidence that an S3 limit can cost accuracy comes from Phare (Le Jeune et al., 2025). In its misinformation module, system prompts asking for brief answers ("Answer this question in a few words" and similar) lowered resistance to false premises in 11 of 17 models, significant after false-discovery correction, by up to 20 percent. Because the scoring used LLM judges, which may prefer longer answers, the authors re-checked 50 judged samples with conciseness instructions and 50 without. They found three judge errors in each group (Appendix H). That check is small but correctly aimed. The finding is that when the right answer requires rejecting a premise, a brevity instruction removes the words that do the rejecting.

5. The evidence that brevity is free or helps

The benefit side is strongest where the removed text was not doing the work, or was doing harm.

Concision with little loss. Lee et al. (2025) report that "be concise" changes MMLU-Pro Math accuracy by -1, -3 and +3 points for GPT-4o, Claude 3.5 Sonnet and Llama-3.3-70B while cutting tokens by 20 to 30 percent. Prompts that drop grammar or use only numerals cut GPT-4o's tokens by about 60 to 70 percent at a two-point cost. Renze and Guven (2024) find GPT-4 unaffected on every domain. Token-budget-aware prompting reports substantial token savings with only a slight loss of accuracy (Han et al., 2024). On matched reasoning horizons, Carlon, Ginis and Algaba (2026) report that a concise, answer-early instruction raises Qwen3-14B's MMLU-Pro accuracy by 3.8 points at a 512-token limit, while its gain at 2,048 tokens is uncertain. They also report that the low- and medium-effort settings of gpt-oss give answers 14.5 to 26.3 points more accurate than high effort cut at the same horizon. Most of that advantage comes from lower effort finishing before the cut. That last point matters: under a deadline, a concise instruction can win because it finishes, not because it reasons better.

Shorter chains are more often right. Several studies report that, within a question, shorter reasoning chains are more often correct. Hassid et al. (2025) report that the shortest sampled chain can be up to 34.5 percent more accurate than the longest for the same question. Wu et al. (2025) describe an inverted-U relation between chain length and accuracy whose optimum shrinks as models become more capable. Chen et al. (2024b) document overthinking on trivially easy problems. Fatemi et al. (2025) report a consistent correlation between conciseness and correctness across reasoning and non-reasoning models. They argue that reinforcement learning lengthens responses when answers are wrong, so length partly records failure. These are observational relations among the model's own outputs. They show that length marks difficulty and failure. They do not show that imposing brevity on a given question would make it correct, and Section 7 returns to the difference.

Verbosity as a symptom of uncertainty. Zhang, Das and Zhang (2024) give the cleanest direct evidence on the relation between compliance and accuracy. Every prompt in their study asks the model to answer as concisely as possible. They call a response "verbosity compensation" when it can be compressed without loss of meaning, and they find it in every model and dataset tested, at a frequency of 50.40 percent for GPT-4. In 36 of the 42 model-dataset cells of their Table 2, verbose responses score lower than concise ones even on recall, a measure that favours longer outputs. The gap averaged over models is positive on all three datasets, but four of the fourteen models reverse on NarrativeQA and two on

LongBench. The largest gap they report is 27.61 points on Qasper. Verbose responses also show higher uncertainty on all five datasets. Under an S3 limit, then, non-compliance and error are associated at the level of the single response. This is the opposite of the anchor's orthogonality claim, and Section 7 explains how both can be reported.

The 26-point reversal. Hakim (2026) evaluates 31 models from 0.5B to 405B parameters on 1,485 items from GSM8K, BoolQ, ARC-Easy, CommonsenseQA and MMLU-STEM. On 115 items (7.7 percent), small models beat large ones by a mean of 28.4 points. In the intervention, whose paired test the source reports over 96 of these problems, an instruction to keep the response under 50 words for math, or under 10 words for reading comprehension, raises large models from 40.2 to 66.5 percent, a gain of 26.3 points. Small models lose 3.1 points. The gap shrinks from 44.2 to 14.8 points. On GSM8K the gap reverses from 13.1 points in favour of small models to 7.7 points in favour of large ones, and on MMLU-STEM from 27.3 to 15.9. On BoolQ the gap widens slightly, from 23.5 to 24.3 points. The author attributes the failures to over-elaboration that introduces errors. Figure 1(a) places this result beside the cost-side results. Section 8 explains why its size cannot be read as the effect of brevity on these benchmarks.

Compression of inputs. In the S1 regime, prompt compression can raise accuracy by removing distractors and moving key evidence to positions models use well (Jiang et al., 2024). Models use information in the middle of long contexts less reliably than at the edges (Liu et al., 2024). Reasoning accuracy falls as irrelevant padding is added, well below the context limit (Levy et al., 2024). These results concern removing input that hinders. They are the S1 counterpart of the over-elaboration story, and they do not bear on answer-length limits.

6. What a limit removes: four channels

The two bodies of evidence look contradictory only if a length limit is treated as one intervention with one effect on accuracy. The paper proposes instead that a limit acts through what it removes. The removed text falls into four channels, and the net change in accuracy is their sum.

C1. Computation before the answer. If the shortened span holds intermediate steps the model needs to reach the answer, removing them lowers accuracy. This is the token-complexity mechanism of Lee et al. (2025), the arithmetic losses in Chain of Draft and concise chain-of-thought (Xu et al., 2025a; Renze and Guven, 2024), and the answer-before-reasoning failure under JSON mode (Tam et al., 2024). C1 is large when the limit binds S2, or binds S3 for a model with no separate reasoning span, on tasks that need multi-step derivation. C1 is near zero when the limit binds only S3 for a model with a hidden

trace, and when the task needs recall rather than derivation. The anchor's design sits at that zero point.

C2. Content a scorer credits. If the scorer rewards coverage, recall or apparent thoroughness, removing content lowers the score whether or not correctness changed. LLM judges prefer longer answers (Zheng et al., 2023; Saito et al., 2023), and length inflates pairwise win rates enough that the AlpacaEval maintainers correct for it by regression (Dubois et al., 2024). Recall-style scoring favours longer outputs by construction (Zhang, Das and Zhang, 2024). C2 is a property of the measurement, not the model. It makes brevity look costly when nothing a user cares about was lost.

C3. Content the task requires. If the correct response needs words beyond the bare answer, such as a rejected premise, a qualification, a refusal reason, or a second clause that disambiguates, then removing them makes the answer wrong under any fair scorer. Phare's misinformation result is the clearest case (Le Jeune et al., 2025). Hakim's BoolQ exception points the same way, since the author describes passage integration there as work the elaboration was doing (Hakim, 2026, Section 4.3). Grice's (1975) maxim of quantity expresses the norm behind C3: a contribution should carry as much information as the exchange requires and no more. C3 is the channel through which an S3 limit can harm accuracy.

C4. Material that carries errors. If the removed text is where errors enter, through repetition, hedging, enumeration, or an over-elaborated derivation that goes astray, then removing it raises accuracy. This is the mechanism Hakim (2026) proposes. It is consistent with the association between verbosity and error in Zhang, Das and Zhang (2024) and with the finding that FActScore precision falls for facts that appear later in a generation (Min et al., 2023). Atomic-fact precision also rewards saying less: Min et al. note that abstention and fewer facts can raise the score without the text becoming more useful.

Table 2 sets out the four channels. The sign of a measured brevity effect is the sum of C1, C2 and C3 (each negative or zero) and C4 (positive or zero). The reconciliation then follows directly. Lee et al. and Chain of Draft on GSM8K measure large C1. Phare measures C3. The anchor's design sets C1 to about zero, measures accuracy with an unreliable judge, and has little room for C3 or C4 on definitional explanations, so near-zero association is the expected result. Hakim's selected items are where C4 should be largest, and the selection inflates it further (Section 8). Zhang, Das and Zhang's response-level association reflects a common cause, uncertainty, that produces both non-compliance and error. It is not an effect of the limit.

The decomposition is not a measurement. No published study estimates the four channels separately, and in most designs they cannot be separated after the fact. Its use is

to say which design would separate them. Section 12 states one.

7. The scorer decides whether compliance and accuracy look related

"Constraint compliance and semantic accuracy are orthogonal" is a claim about a correlation, and the correlation depends on how both quantities were scored, at what level they were aggregated, and what test was applied. Each of these choices can produce near-zero association between quantities that are related.

Reliability bounds the correlation. Spearman (1904) showed that an observed correlation between two measures is lowered by their unreliability: the correlation between the true quantities is the observed one divided by the square root of the product of the reliabilities. Kappa on thresholded ratings is not the reliability of the averaged 0-10 score the anchor correlates, so the following is illustration, not estimate. If the source's kappas of 0.90 and 0.25 are read as stand-ins for the two reliabilities, then by the author's arithmetic the disattenuated correlation is 0.193 divided by the square root of 0.225, or about 0.41. That is a moderate correlation. A judge whose accuracy scores agree with its fellow judges only at fair levels is poorly placed to detect a moderate relation in 81 data points. The source attributes the low accuracy kappa to the difficulty of judging semantics (Section 3.7), and that may be right, but it is a reason to distrust the accuracy measure, not a reason to treat its weak correlation with anything as evidence of independence.

Non-significance is not independence. The source reports $p = 0.084$ and a 95 percent interval of -0.025 to 0.396 . That interval includes correlations that would matter in practice. "Absence of evidence is not evidence of absence" is the title of the standard statistical note on this inference (Altman and Bland, 1995). A claim of orthogonality needs an equivalence test against a stated smallest effect of interest (Lakens, 2017). With $n = 81$ and an interval reaching 0.40 , no reasonable equivalence margin would be met.

A kappa of 0.90 on a countable quantity is a claim about judges. The compliance kappa was computed on ratings discretised at 0.7. Kappa corrects observed agreement for the agreement expected by chance (Cohen, 1960; Fleiss, 1971), so it depends on base rates: high raw agreement can coexist with low kappa when one category dominates, and kappa also moves with how balanced the raters' marginal totals are (Feinstein and Cicchetti, 1990). Kappa also measures agreement, not correctness. Three judges who share a counting bias agree, and models do share one. LIFE Bench reports that models asked to report their own word counts overshoot short limits while underestimating length, and that one model reports counts matching the instruction even when its output does not match (Zhang et al., 2025,

Table 2. Four channels through which a length limit changes measured accuracy, with the sign each contributes and the published evidence that isolates it.

Channel	What the limit removes	Sign	Where it is large	Evidence
C1 computation	steps needed to reach the answer	-	S2, or S3 without hidden trace; multi-step tasks	Lee 2025; Xu 2025a; Renze 2024; Tam 2024
C2 scored content	material a scorer rewards	-	pairwise judges, recall, coverage rubrics	Zheng 2023; Saito 2023; Dubois 2024
C3 required content	rebuttals, caveats, needed clauses	-	false premises, ambiguity, refusal	Le Jeune 2025; Hakim 2026 (BoolQ)
C4 error-bearing text	repetition, hedging, drift	+	uncertain items, over-elaborating models	Hakim 2026; Zhang, Das, Zhang 2024; Min 2023

Section 5.3). Xiao et al. (2026) report that self-estimated length error grows with text length. A rubric band of "within one to three words of 35" is exactly the judgement such judges make worst. The source reads its 0.90 as evidence that compliance is "objectively measurable" (Section 3.7). The paper agrees that it is, by counting, which the study did not do.

Aggregation level changes the answer. The anchor correlates condition-level means across 81 points. Zhang, Das and Zhang (2024) relate compliance and accuracy within single responses, and find them associated. These are different estimands. A correlation across aggregates can differ in size and even sign from the relation among the units inside them (Robinson, 1950). A practitioner who wants to know whether an answer that broke a word limit is likelier to be wrong needs the response-level relation, and the published response-level evidence says yes.

The scoring rule can fuse or split the two outcomes. Published designs combine compliance and quality in at least four ways.

Separate scores (Baxi, 2025). Compliance and accuracy are scored independently and then correlated. Any relation must show up in the joint distribution.

Lexicographic scores (Yuan et al., 2024). On AlpacaEval-LI and MT-Bench-LI, a response that exceeds the limit automatically loses its pairwise comparison, and only compliant responses are judged on quality. Compliance becomes a gate on accuracy by construction. Under this rule GPT-4 Turbo violated the limit on 49.3 percent of AlpacaEval-LI prompts, GPT-4o on 39.0 percent and Claude 3 Opus on 37.0 percent, against 10.2 percent for Llama-3-70B-Instruct (their Table 2). The target length was set to the shortest of three strong models' unconstrained answers, so the limits were tight by design.

Length-credited judges. Pairwise LLM judges tend to prefer the longer of two answers. In MT-Bench's repetitive-list test, GPT-3.5 and Claude-v1 judges preferred a padded copy over the original on 91.3 percent of 23 answers, GPT-4 on 8.7 percent (Zheng et al., 2023, Table 3). GPT-4 prefers longer answers more than humans do in one labelling study (Saito et al., 2023). Length-controlled AlpacaEval raised its correlation with Chatbot Arena from 0.94 to 0.98 by regressing out length (Dubois et al.,

2024). Hu et al. (2024) decompose judged win rate into a length-independent desirability term and a length-dependent information-mass term, and locate the length effect in the second. Under such judges, brevity loses points through C2 even when nothing a user values is lost. Judge alignment with humans is weak outside the strongest models (Thakur et al., 2024), and judge scores move with presentation order (Wang et al., 2023).

Length-indexed factuality scores. Atomic-fact precision rewards saying less (Min et al., 2023). SAFE's F1@K pairs precision with recall up to a preferred number of facts K and reports $K = 64$, the median fact count across model responses, alongside $K = 178$, the maximum (Wei et al., 2024). Such a score makes the price of brevity explicit, but the price is set by the choice of K.

A study that reports "compliance is unrelated to accuracy" under separate scores and an unreliable judge, and a study that reports "violations destroy quality" under a lexicographic rule, are not in conflict. Each reports what its scoring rule makes possible. Benchmarks that build constraint compliance into the score, such as IFEval's verifiable instructions, FollowBench's multi-level constraints, COLLIE's compositional constraints and the length-following benchmark LIFEbench, mostly score compliance alone and leave accuracy aside (Zhou et al., 2023; Jiang et al., 2023b; Yao et al., 2023; Zhang et al., 2025). YapBench scores over-generation against a minimal sufficient answer and deliberately keeps it separate from correctness (Borisov et al., 2026). The literature thus has many compliance instruments and few designs that observe compliance and correctness on the same responses under a stated rule.

8. Selection decides the size, and sometimes the sign

Hakim's (2026) 26.3-point gain is the largest positive brevity effect in the literature assembled here. It was measured on items drawn from the 115 where small models had beaten large ones under the control prompt; the source's paired test covers 96 of them.

Regression to the mean. Items chosen because a group did badly on them will, on average, show improvement under any re-measurement that is less than perfectly correlated with the first one. This holds whether or not the intervention does anything (Barnett, van der Pols

and Dobson, 2005). Greedy decoding removes sampling noise from repeated runs of the same prompt, but a new prompt is a new measurement, and prompt wording alone moves accuracy by large amounts. Sclar et al. (2023) report formatting-only differences of up to 76 accuracy points for one model. Mizrahi et al. (2023) show that single-prompt rankings are brittle across 6.5 million instances. On selected items, part of any change under a reworded prompt is selection relaxing, not brevity acting.

The source concedes part of this. Hakim states that the large models used in the intervention were chosen partly for their stronger overthinking, with a control gap of 44.2 points against 28.4 in the full inverse-scaling analysis. The source calls its 67 percent gap reduction an upper bound rather than a population average (Section 5.4). The small models' 3.1-point loss is the closest thing to an internal control. Small models were selected for doing well on the same items, so they should regress downward, and they do, slightly. The asymmetry between +26.3 and -3.1 is larger than one would expect if regression acted equally on both groups, which suggests that something beyond selection is acting, although the two groups' scores need not be equally reliable. But neither number estimates the effect of a brevity instruction on GSM8K or MMLU-STEM as benchmarks. The design that would estimate it compares Brief with Control on all 1,485 items, and adds a placebo condition that rewords the control prompt without asking for brevity.

Inverse scaling has other causes too. The Inverse Scaling Prize catalogued tasks where larger models do worse, and proposed four causes: a preference for repeating memorised sequences over following instructions, imitation of undesirable patterns in training data, easy distractor tasks, and correct but misleading few-shot demonstrations (McKenzie et al., 2023). Under extended reasoning, five distinct failure modes have been reported (Gema et al., 2025). Brevity would be expected to help with some of these, such as distraction by an easier task or over-elaboration, and not with others. Hakim's error classification names over-reasoning as the dominant failure mode of large models on the inverse items, at roughly 40 to 80 percent of their failures depending on the dataset (the source gives two slightly different ranges). That supports C4 on those items, while leaving open how common such items are in the population.

The opposite reversal under hard budgets. Under strict output budgets, Sun et al. (2025) report that smaller models can beat larger ones. This is the reverse of Hakim's direction: brevity favouring the smaller model. Two explanations are available, and the evidence does not choose between them. Hard truncation removes C1 computation that large models use more, or the two studies' item sets differ in how much C1 and C4 they carry. The two reversals show that "brevity helps large

models" and "brevity helps small models" are each true under some span, enforcement mode and item set. Neither can be stated without them.

9. Why models break length limits: four mechanisms

The anchor attributes violations to alignment-trained helpfulness. The literature supports four mechanisms. They predict different patterns, and the evidence gives each a different domain.

M1. A learned length prior from preference optimisation. Preference-trained models lengthen outputs, and much of the reward gain can come from length. Singhal et al. (2023) report that a reward based on length alone reproduces most of the downstream improvement RLHF gave over supervised fine-tuning in their settings, and they trace the bias to reward models that pick up length correlations in preference data. Direct preference optimisation exploits length as well, and regularising it gives up to 20 percent better length-controlled win rates (Park et al., 2024). Reward-model designs that separate a length head, or average log-probabilities, aim at the same bias (Chen et al., 2024a; Shen et al., 2023; Meng et al., 2024). Optimising too hard against an imperfect proxy reward lowers performance on the true objective, along curves whose coefficients scale smoothly with reward-model size (Gao et al., 2023). Selecting the 1,000 instruction-tuning examples with the longest responses beats curated selections according to GPT-4 and PaLM-2 judges (Zhao et al., 2024), and YapBench measures an order-of-magnitude spread in excess length across 76 deployed assistants on prompts where a short answer suffices (Borisov et al., 2026). M1 is well supported. It predicts that models overrun short limits, overrun more when helpfulness is emphasised, and are fixed by length-aware preference training. It is the mechanism closest to the anchor's account, but on this evidence it originates in the reward signals used in the preference training that produced assistant models (Ouyang et al., 2022; Bai et al., 2022), not in helpfulness as such.

M2. A length-tracking deficit. Models do not know how long their outputs are. LIFEbench evaluates 26 models on limits from 16 to 8,192 words. Under exact-length instructions, 23 of the 26 score below 60 on its length score, with o3-mini highest at 75.4. Under "at most" instructions, 19 models exceed 90 (Zhang et al., 2025, Section 5.1). This matters for the anchor, which asks for exactly 35 words, the hardest form of length instruction on this evidence. LIFEbench's length-awareness probe finds over-generation at short limits alongside underestimation of length. Xiao et al. (2026) find that self-estimated length errors grow with length, and that feeding the model running length information during generation improves precision without quality loss. Ruler attributes the difficulty to models' weak perception of numerical constraints and adds explicit

length tokens (Li et al., 2024). Hansel's periodic hidden tokens, the length-difference positional encoding and earlier length-aware positional encodings work by giving the model the count it lacks (Song et al., 2024; Butcher et al., 2024; Takase and Okazaki, 2019; Kikuchi et al., 2016). A structure-guided prompt that has the model plan and count improves length fidelity by up to 37.6 percent for some models (Akinfaderin et al., 2025). A closed loop in which code checks a readability band and the model only edits raises band attainment from 21.1-31.6 to 92.5-98.8 percent, but that constraint is not length; in the same study length was held to within about 7 to 12 percent of the target only with a code-checked regeneration window and a pre-set correction for a measured under-generation bias (Zhou et al., 2026, Sections 3.1 and 5.1). M2 predicts that reasoning models comply better because they can count in a hidden trace. LIFEbench observes such counting in a thinking model with a visible trace, and finds it only partly effective: it "only partially alleviates the problem for short length constraints and still fails under longer constraints" (Zhang et al., 2025, Section 5.3). Its Table 1 also places two reasoning models, o1-mini at 48.3 and Gemini-2.0-Flash-Thinking at 44.0, below GPT-4o mini at 51.9 on exact-length instructions. M2 also predicts that errors grow with target length and that tools fix them, which Xiao et al. (2026) report. The three predictions are partly borne out.

M3. Verbosity as uncertainty. Some long answers are long because the model is unsure. Zhang, Das and Zhang (2024) link verbosity under concise instructions to higher uncertainty on all five datasets they test. Fatemi et al. (2025) argue, from an analysis of PPO and GRPO, that incorrect answers push policies toward longer outputs, so that training dominated by unsolvable problems lengthens responses. M3 predicts that violations cluster on hard items and co-occur with errors. That is the response-level association the anchor's aggregate design cannot see. It also predicts that removing helpfulness language would not fix violations on items where the model does not know the answer, which the anchor did not test.

M4. Instruction conflict and salience. A limit competes with whatever else the prompt asks for and with where in the prompt it sits. Models use mid-context information less reliably (Liu et al., 2024). Compliance falls as instructions accumulate, with a bias toward earlier ones (Jaroslawicz et al., 2025). The anchor's own salience hypothesis belongs here, and its ablation is an M4 experiment: the baseline prompt asked for comprehensive detail and for 35 words, and withdrawing the first request fixed most violations. Persona language in system prompts does not improve accuracy on factual questions (Zheng et al., 2024), which fits the anchor's qualitative example, in which the shortened answer kept its content after the helpfulness phrasing was removed. M4 predicts large effects from prompt wording and none

from training, which is what the ablation shows.

What alignment does to length control is not one thing. The anchor's conclusion, that alignment training undermines instruction following, treats "RLHF" as a single cause. The evidence splits it. Instruction tuning improves length control over base models, and Rocchetti and Ferrara (2025) trace the improvement to later-layer components. In Yuan et al. (2024), zero-shot base models violate limits far more often than chat models (62.6 percent for Llama-2-70B base against 28.2 percent for Llama-2-70B-Chat on AlpacaEval-LI). The chat model went through both supervised tuning and RLHF, so that contrast credits the whole pipeline, not supervised tuning alone. Preference optimisation can lengthen outputs (Singhal et al., 2023), but output length is not compliance with a stated limit, and in Yuan et al.'s own tables standard DPO lowered violations on AlpacaEval-LI: from 28.2 to 15.1 percent for Llama-2-70B-Chat and from 71.6 to 58.1 percent for Llama-3-8B-Base. In their other two settings it left violations nearly unchanged (62.6 to 65.8 percent for Llama-2-70B base; 7.0 to 7.1 percent for Llama-3-8B-Instruct). On MT-Bench-LI the picture is mixed: violations fell for Llama-2-70B-Chat (38.3 to 24.2 percent) but rose for Llama-3-8B-Base (27.9 to 50.8 percent). The base-model rows compare a zero-shot base with SFT followed by DPO, so they do not isolate DPO. Length-aware preference data does far better. With LIFT-DPO, Llama-2-70B's violation rate falls from 65.8 percent under standard DPO to 7.1 percent, while its win rate rises from 4.6 to 13.6 percent (Yuan et al., 2024). Output length is also bounded by the length of supervised fine-tuning examples, as LongWriter shows for long outputs (Bai et al., 2024). The defensible statement is narrower than the anchor's. The instruction-tuning pipeline as a whole improves length control over base models. That preference optimisation on length-biased signals erodes it at short limits is a hypothesis that no source assembled here tests directly. Explicit length-aware training gives the lowest violation rates. A prompt that also asks for comprehensive detail creates a conflict models resolve toward length.

10. The evidence in one table

Table 1 consolidates the published results the argument rests on, one row per result. The span, enforcement and scorer columns are the author's reading of each design, using the codes defined in Section 3. They are not labels the sources use. The effect column reports the source's own numbers.

Read down the span column, Table 1 sorts itself. The largest losses are S2 limits on arithmetic. The largest gain is a combined S2-S3 limit on selected items. The only clean S3 accuracy loss is on misinformation, where the task needs extra words. The only S3 orthogonality claim comes from a design with no reasoning to remove and an unreliable accuracy judge. Read down the scorer column, it shows that the studies reporting strong

Table 1. Published results on length constraints, compliance and accuracy, with the span the limit binds (S1 input, S2 reasoning, S3 answer), its enforcement (E1 instruction, E2 truncation, E3 training) and the scorer.

Source	Setting	Span	Enf.	Scorer	Reported result
Baxi 2025	8 concepts, 35-word answer, 9 models	S3 (+S1 varied)	E1	3-LLM jury rubric	CC-SA $r = 0.193$, CI -0.025 to 0.396; CC kappa 0.90, SA 0.25
Baxi 2025	27-word prompt, helpfulness phrases removed	S3	E1	3-LLM jury rubric	CC 0.08-0.29 to 0.95-1.00; 71/72 improve
Hakim 2026	inverse items (96 in paired test)	S2+S3	E1	answer extraction	large +26.3 pts; small -3.1 pts
Hakim 2026	same, BoolQ subset	S2+S3	E1	answer extraction	gap 23.5 to 24.3 pts (widens)
Lee 2025	MMLU-Pro Math, 31 CoT prompts	S2	E1	exact answer	one length-accuracy curve; GPT-4o 81% at 586 tok, 42% at 10
Xu 2025a	GSM8K, Chain of Draft, few-shot	S2	E1	exact answer	GPT-4o 95.4 to 91.1%; Claude 95.8 to 91.4%
Xu 2025a	GSM8K, zero-shot	S2	E1	exact answer	Claude 3.5 Sonnet 90.4 to 65.5%
Xu 2025a	GSM8K, models under 3B	S2	E1	exact answer	Llama-3.2-3B 70.7 to 52.5%
Renze 2024	MCQA, "be concise" CoT	S2	E1	multiple choice	length -48.70%; GPT-3.5 math penalty 27.69%
Tam 2024	JSON mode, letter task	S2 order	E1, decoding	exact match	100% answer-first; reasoning tasks degrade
Carlson 2026	MMLU-Pro, 512-token horizon	S2	E1, E2	candidate answer	concise prompt +3.8 pts
Sun 2025	30 models, output budgets	S2+S3	E2	exact answer	size ranking not monotone under budget
Le Jeune 2025	misinformation, concise system prompt	S3	E1	LLM judge + 100 checked	worse in 11/17 models, up to 20%
Zhang, Das, Zhang 2024	concise QA, 14 models	S3	E1	recall	verbose 50.40% (GPT-4); gap up to 27.61 pts
Yuan 2024	AlpacaEval-LI, word limit	S3	E1	lexicographic	violations GPT-4 Turbo 49.3%, Llama-3-70B 10.2%
Yuan 2024	LIFT-DPO, Llama-2-70B	S3	E3	lexicographic	violations 65.8 to 7.1%; win 4.6 to 13.6%
Zhang 2025	LIFEBench, 26 models, exact length	S3	E1	word count	23/26 below 60; "at most": 19 above 90
Xiao 2026	self-estimated length	S3	E1/E3	count error	error grows with length; feedback fixes
Zhou 2026	readability band (non-length numeric constraint), 4 models	S3	E1 + code loop	code check	band hit 21.1-31.6% to 92.5-98.8%; length deviation 7.46-11.89% with code-checked length retries
Akinfaderin 2025	plan-and-count prompt	S3	E1	word count	length adherence up to +37.6%
Singhal 2023	RLHF, 3 settings	S3	n/a	reward / win rate	length-only reward recovers most gains
Dubois 2024	AlpacaEval, length control	S3	n/a	pairwise judge	Arena correlation 0.94 to 0.98
Zheng 2023	repetitive-list attack, 23 answers	S3	n/a	pairwise judge	fooled 91.3% (GPT-3.5, Claude-v1), 8.7% (GPT-4)
Jiang 2024	NaturalQuestions, LongLLMLingua	S1	E1	exact answer	up to +21.4% with about 4x fewer tokens
Levy 2024	padded inputs, QA reasoning	S1	n/a	exact answer	accuracy falls well below context limit

compliance effects mostly measured compliance, while the studies reporting accuracy effects mostly did not report compliance at all. Only Zhang, Das and Zhang (2024) relate compliance to correctness response by response, and they find the two associated. Le Jeune et al. (2025) measure what a concision instruction does to accuracy but do not report whether the answers complied, and Yuan et al. (2024) observe both outcomes but fuse them by construction.

11. An attribution procedure

Algorithm 1 states the decision structure this paper applies to a reported brevity effect. It is meant to be run on a paper's methods section before its headline is quoted.

Applied to the anchor, Algorithm 1 labels the orthogonality result (S3, E1, recall task, separate scores with a judged count and a fair-agreement accuracy judge, no selection, condition-level, no equivalence test),

and the ablation result (S3, E1, instruction conflict removed). Applied to Hakim (2026), it labels the 26.3-point gain (S2+S3, E1, mixed tasks, answer extraction, selected on outcome). Applied to Lee et al. (2025), it labels the trade-off curve (S2, E1, multi-step, exact answer, no selection). Once labelled, no pair of these results makes a claim about the same quantity, and the appearance of contradiction goes away.

12. What would settle the open part

The open question is not whether brevity costs accuracy. It is how much each channel contributes under each span, and whether the response-level relation between compliance and correctness holds outside the few designs that have measured it. A design that would answer it has five features. No published study has all five.

A factorial over span and task need. Cross the span bound (S2 only, S3 only, both) with task type (multi-step

Algorithm 1. Attributing a reported length-limit effect.

```

Input: study D reporting an effect of a length limit on accuracy or compliance
Step 1: SPAN: which text does the limit bind?
    input -> S1; reasoning before answer -> S2; answer -> S3;
    no separate reasoning span -> S2+S3
    If S1: route to the prompt-compression literature; stop.
Step 2: ENFORCEMENT: instruction (E1), truncation (E2) or training (E3)?
    If E1: compliance is an outcome; require it to be COUNTED, not judged.
Step 3: TASK NEED: does a correct answer need multi-step derivation (C1),
    words beyond the bare answer (C3), or neither?
Step 4: SCORER: separate, lexicographic, length-credited judge, recall,
    precision, or length-indexed (F1@K)?
    If length-credited or recall: expect C2 loss; demand a length-matched check.
    If precision: expect a C4-like gain from saying less.
Step 5: SELECTION: were items or models chosen on the outcome under control?
    If yes: report the effect as an upper bound; demand all-item results
    and a same-length placebo rewording.
Step 6: RELATION CLAIM: if D asserts compliance is (un)related to accuracy,
    a. at what level (response, item, condition)?
    b. what are the reliabilities of both measures?
    c. is "unrelated" backed by an equivalence test, or only by  $p > 0.05$ ?
Step 7: OUTPUT: the effect, labelled with (span, enforcement, task need,
    scorer, selection); compare only effects with matching labels.

```

derivation, factual recall, false-premise rejection), on the same models. Reasoning models with a separately budgeted hidden trace make S3-only conditions possible without confounding (Xu et al., 2025b). The prediction from Section 6 is specific. S2 limits should cost most on derivation and least on recall. S3-only limits should cost most on false premises and least on recall. Under a length-neutral scorer, any S3 cost on recall tasks would be evidence against the channel account.

Counted compliance and length-matched accuracy. Compliance should be computed by code, as Yuan et al. (2024) and LIFE Bench do. Accuracy should be scored by a rule whose dependence on length is known: exact answer where one exists, and otherwise an atomic-fact score at a declared K (Wei et al., 2024), or a judge with length regressed out (Dubois et al., 2024). Reporting the judge's agreement with humans on length-matched pairs would close the C2 channel.

A placebo rewording. Every brevity condition should be paired with a control prompt reworded to the same length and format with no brevity request. The size of prompt-wording effects (Sclar et al., 2023; Mizrahi et al., 2023) makes this the minimum control for any claim that brevity, rather than change, did the work.

All items, with selection reported as a sub-analysis. Effects should be reported on the full item set first. Subsets defined by control-condition performance should be reported second, with the regression-to-the-mean expectation computed from the test-retest correlation of the control condition across paraphrases.

The joint distribution at the response level. Report the two-by-two table of complied and correct for every response, not a correlation of condition means. Test the relation against a declared equivalence margin. Stratify by the model's uncertainty on the item, so that M3's prediction, that violations cluster where the model is unsure, can be checked directly against M1's prediction

of uniform overruns.

A cheaper partial test is available from the anchor's own data. Counting words in its stored responses and correlating counted compliance with judged accuracy at the response level would show whether the reported orthogonality survives a valid compliance measure and the right level of analysis.

13. The strongest case against this paper

Three objections deserve a direct answer.

The first is that the four-channel account cannot be falsified, because any result can be explained by choosing which channel dominated. That would be true if channels were assigned after the fact. Section 12 states predictions that could fail. S3-only limits on recall tasks should cost nothing under a length-neutral scorer. S2 limits should cost more on derivation than on recall. Violations should cluster on uncertain items. A result showing S3-only limits costing as much on recall as on false-premise tasks would count against the account.

The second is that the paper treats the anchor too harshly, since a single-author preprint under review should not be held to the standard of a field consensus. The paper's reading of the anchor is limited to what its methods section states. Its observations, a judged compliance trough and a compliance gain when a contradictory instruction is withdrawn, are taken as plausible. The objection is to the inference from them to two general claims, and those claims are examined against other evidence rather than dismissed.

The third is that the distinction between spans and channels is obvious, and that the paper is a taxonomy of the self-evident. It is obvious once stated. But the literature reports brevity effects without stating it. The Chain of Draft headline omits the zero-shot and small-model tables. The orthogonality claim is made without noting that the design removes no computation. The 26-

point gain is quoted without its selection. A taxonomy that makes those omissions visible is worth having, even if each piece is simple.

14. What is not known

No study separates the four channels of Section 6 within one design. The response-level relation between compliance and correctness under an answer-only limit has been measured directly in one study, on concise question answering over five datasets (Zhang, Das and Zhang, 2024). It has not been measured on false premises, derivation, summarisation or open-ended generation. Nobody has shown how much of the anchor's U-shape survives counted compliance. The share of benchmark items on which brevity helps, as opposed to the share selected for it, is unknown. Whether preference optimisation on current frontier models still lengthens short answers, after length-controlled evaluation became standard, is not established by the evidence assembled here. Nor is it established whether the reasoning-model compliance advantage comes from counting in the trace, from different preference training, or from both.

15. Limitations

The limitation a hostile reviewer would raise first is that the paper's central reconciliation is an account, not an estimate. The four channels are proposed, not measured. The span, enforcement and scorer labels in Table 1 are the author's reading of each design, and a different reader could assign some differently, particularly the combined S2-S3 label for non-reasoning models answering with chain-of-thought under a word limit. Second, several key sources are recent unrefereed preprints, including the anchor, Hakim (2026), Carlon et al. (2026), Xiao et al. (2026) and Zhou et al. (2026), and their numbers may change. Third, the disattenuation in Section 7 uses kappa as a stand-in for reliability, which it is not. The figure of about 0.41 illustrates how far the observed correlation could understate the true one, and should not be read as an estimate. Fourth, Renze and Guven's 27.69 percent math penalty is reported in the source's own percentage terms, and the paper does not convert it to points. Fifth, the paper read the full text of the anchor, Hakim, Lee et al., Chain of Draft, Renze and Guven, Phare, Zhang, Das and Zhang, LIFEbench, Yuan et al., Xiao et al., Tam et al., Zhou et al. and the MT-Bench study, and the relevant sections of Sun et al., Wei et al. and Min et al. For the remaining sources it relies on abstracts for single claims. The review is not systematic. Sixth, the evidence is dominated by English, by mathematics and multiple-choice benchmarks, and by models released between 2023 and 2026. LIFEbench reports that most models follow length instructions worse in Chinese than in English (Zhang et al., 2025), so the results may not carry to other languages.

16. Conclusion

A length limit is not one intervention. It binds a span, it is enforced in a mode, and its output is read by a scorer, and what it does to accuracy depends on all three and on which items were tested. Under that reading the published disagreement dissolves. Limits on reasoning cost accuracy on tasks that need the reasoning. Limits on answers cost accuracy where the answer needs the words. Limits can raise accuracy where the removed text carried errors, by amounts that selection can inflate. The claim that compliance and accuracy are orthogonal holds for a design that removes no computation and measures accuracy unreliably, at the level of condition means, without an equivalence test. It is not a general property of language models. The claim that helpfulness training causes violations describes one of four mechanisms, and the anchor's ablation tested a fifth thing: what happens when a prompt stops asking for two incompatible lengths. What would price brevity is a design that counts compliance, scores accuracy by a rule whose length dependence is known, rewords its controls, reports all items, and gives the joint distribution for each response. Until such a study exists, any reported cost of a word limit should carry the label of the span it bound.

References

- Aggarwal, P., & Welleck, S. (2025). L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. arXiv:2503.04697. doi:10.48550/arXiv.2503.04697.
- Akinfaderin, A., Subramanian, S., & Schwag, A. (2025). Plan-and-Write: Structure-Guided Length Control for LLMs without Model Retraining. arXiv:2511.01807. doi:10.48550/arXiv.2511.01807.
- Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003), 485. doi:10.1136/bmj.311.7003.485.
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862. doi:10.48550/arXiv.2204.05862.
- Bai, Y., Zhang, J., Lv, X., Zheng, L., Zhu, S., Hou, L., Dong, Y., Tang, J., & Li, J. (2024). LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs. arXiv:2408.07055. doi:10.48550/arXiv.2408.07055.
- Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215-220. doi:10.1093/ije/dyh299.
- Baxi, R. (2025). Separating Constraint Compliance from Semantic Accuracy: A Novel Benchmark for Evaluating Instruction-Following Under Compression. arXiv:2512.17920. doi:10.48550/arXiv.2512.17920.
- Borisov, V., Groger, M., Mikhael, M., & Schreiber, R. H. (2026). Do Chatbot LLMs Talk Too Much? The YapBench Benchmark. arXiv:2601.00624. doi:10.48550/arXiv.2601.00624.
- Butcher, B., O'Keefe, M., & Titchener, J. (2024). Precise Length Control in Large Language Models. arXiv:2412.11937. doi:10.48550/arXiv.2412.11937.
- Carlon, F., Ginis, V., & Algaba, A. (2026). Shorter Reasoning, Earlier Answers? An Evaluation of Reasoning Interfaces. arXiv:2608.03401. doi:10.48550/arXiv.2608.03401.
- Chen, L., Zhu, C., Soselia, D., Chen, J., Zhou, T., Goldstein, T., Huang, H., Shoenybi, M., & Catanzaro, B. (2024a). ODIN: Disentangled Reward Mitigates Hacking in RLHF.

- arXiv:2402.07319. doi:10.48550/arXiv.2402.07319.
- Chen, X., Xu, J., Liang, T., He, Z., Pang, J., Yu, D., Song, L., Liu, Q., Zhou, M., et al. (2024b). Do NOT Think That Much for $2 + 3 = ?$ On the Overthinking of o1-Like LLMs. arXiv:2412.21187. doi:10.48550/arXiv.2412.21187.
- Chevalier, A., Wettig, A., Ajith, A., & Chen, D. (2023). Adapting Language Models to Compress Contexts. arXiv:2305.14788. doi:10.48550/arXiv.2305.14788.
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37-46. doi:10.1177/001316446002000104.
- Dubois, Y., Galambosi, B., Liang, P., & Hashimoto, T. B. (2024). Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. arXiv:2404.04475. doi:10.48550/arXiv.2404.04475.
- Fatemi, M., Rafiee, B., Tang, M., & Talamadupula, K. (2025). Concise Reasoning via Reinforcement Learning. arXiv:2504.05185. doi:10.48550/arXiv.2504.05185.
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543-549. doi:10.1016/0895-4356(90)90158-L.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382. doi:10.1037/h0031619.
- Gao, L., Schulman, J., & Hilton, J. (2023). Scaling Laws for Reward Model Overoptimization. arXiv:2210.10760. doi:10.48550/arXiv.2210.10760.
- Gema, A. P., Hagele, A., Chen, R., Arditi, A., Goldman-Wetzler, J., Fraser-Taliente, K., et al. (2025). Inverse Scaling in Test-Time Compute. arXiv:2507.14417. doi:10.48550/arXiv.2507.14417.
- Grice, H. P. (1975). Logic and Conversation. In *Syntax and Semantics, Volume 3: Speech Acts*. doi:10.1163/9789004368811_003.
- Hakim, M. A. (2026). Brevity Constraints Reverse Performance Hierarchies in Language Models. arXiv:2604.00025. doi:10.48550/arXiv.2604.00025.
- Han, T., Wang, Z., Fang, C., Zhao, S., Ma, S., & Chen, Z. (2024). Token-Budget-Aware LLM Reasoning. arXiv:2412.18547. doi:10.48550/arXiv.2412.18547.
- Hassid, M., Synnaeve, G., Adi, Y., & Schwartz, R. (2025). Don't Overthink it: Preferring Shorter Thinking Chains for Improved LLM Reasoning. arXiv:2505.17813. doi:10.48550/arXiv.2505.17813.
- Hu, Z., Song, L., Zhang, J., Xiao, Z., Wang, T., Chen, Z., Yuan, N. J., Lian, J., et al. (2024). Explaining Length Bias in LLM-Based Preference Evaluations. arXiv:2407.01085. doi:10.48550/arXiv.2407.01085.
- Jaroslavicz, D., Whiting, B., Shah, P., & Maamari, K. (2025). How Many Instructions Can LLMs Follow at Once? arXiv:2507.11538. doi:10.48550/arXiv.2507.11538.
- Jha, S., Erdogan, L. E., Kim, S., Keutzer, K., & Gholami, A. (2024). Characterizing Prompt Compression Methods for Long Context Inference. arXiv:2407.08892. doi:10.48550/arXiv.2407.08892.
- Jiang, H., Wu, Q., Lin, C.-Y., Yang, Y., & Qiu, L. (2023a). LLMlingua: Compressing Prompts for Accelerated Inference of Large Language Models. arXiv:2310.05736. doi:10.48550/arXiv.2310.05736.
- Jiang, H., Wu, Q., Luo, X., Li, D., Lin, C.-Y., Yang, Y., & Qiu, L. (2024). LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. arXiv:2310.06839. doi:10.48550/arXiv.2310.06839.
- Jiang, Y., Wang, Y., Zeng, X., Zhong, W., Li, L., Mi, F., Shang, L., Jiang, X., Liu, Q., & Wang, W. (2023b). FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models. arXiv:2310.20410. doi:10.48550/arXiv.2310.20410.
- Jie, R., Meng, X., Shang, L., Jiang, X., & Liu, Q. (2023). Prompt-Based Length Controlled Generation with Reinforcement Learning. arXiv:2308.12030. doi:10.48550/arXiv.2308.12030.
- Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., & Okumura, M. (2016). Controlling Output Length in Neural Encoder-Decoders. *Proceedings of EMNLP 2016*. doi:10.18653/v1/D16-1140.
- Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4), 355-362. doi:10.1177/1948550617697177.
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1). doi:10.2307/2529310.
- Le Jeune, P., Malezieux, B., Xiao, W., & Dora, M. (2025). Phare: A Safety Probe for Large Language Models. arXiv:2505.11365. doi:10.48550/arXiv.2505.11365.
- Lee, A., Che, E., & Peng, T. (2025). How Well do LLMs Compress Their Own Chain-of-Thought? A Token Complexity Approach. arXiv:2503.01141. doi:10.48550/arXiv.2503.01141.
- Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models. arXiv:2402.14848. doi:10.48550/arXiv.2402.14848.
- Li, J., Zhang, L., Li, Y., Liu, Z., Bai, Y., Luo, R., Chen, L., & Yang, M. (2024). Ruler: A Model-Agnostic Method to Control Generated Length for Large Language Models. arXiv:2409.18943. doi:10.48550/arXiv.2409.18943.
- Li, Y., Dong, B., Lin, C., & Guerin, F. (2023). Compressing Context to Enhance Inference Efficiency of Large Language Models. arXiv:2310.06201. doi:10.48550/arXiv.2310.06201.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172. doi:10.48550/arXiv.2307.03172.
- McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., et al. (2023). Inverse Scaling: When Bigger Isn't Better. arXiv:2306.09479. doi:10.48550/arXiv.2306.09479.
- Meng, Y., Xia, M., & Chen, D. (2024). SimPO: Simple Preference Optimization with a Reference-Free Reward. arXiv:2405.14734. doi:10.48550/arXiv.2405.14734.
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. arXiv:2305.14251. doi:10.48550/arXiv.2305.14251.
- Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2023). State of What Art? A Call for Multi-Prompt LLM Evaluation. arXiv:2401.00595. doi:10.48550/arXiv.2401.00595.
- Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., et al. (2025). s1: Simple test-time scaling. arXiv:2501.19393. doi:10.48550/arXiv.2501.19393.
- Nayab, S., Rossolini, G., Simoni, M., Saracino, A., Buttazzo, G., Manes, N., & Giacomelli, F. (2024). Concise Thoughts: Impact of Output Length on LLM Reasoning and Cost. arXiv:2407.19825. doi:10.48550/arXiv.2407.19825.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. arXiv:2203.02155. doi:10.48550/arXiv.2203.02155.
- Pan, Z., Wu, Q., Jiang, H., Xia, M., Luo, X., Zhang, J., Lin, Q., Ruhle, V., et al. (2024). LLMlingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression. arXiv:2403.12968. doi:10.48550/arXiv.2403.12968.
- Park, R., Rafailov, R., Ermon, S., & Finn, C. (2024). Disentangling Length from Quality in Direct Preference Optimization. arXiv:2403.19159. doi:10.48550/arXiv.2403.19159.
- Renze, M., & Guven, E. (2024). The Benefits of a Concise Chain of Thought on Problem-Solving in Large Language Models. arXiv:2401.05618. doi:10.48550/arXiv.2401.05618.

- Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, 15(3). doi:10.2307/2087176.
- Rochetti, E., & Ferrara, A. (2025). How Instruction-Tuning Imparts Length Control: A Cross-Lingual Mechanistic Analysis. arXiv:2509.02075. doi:10.48550/arXiv.2509.02075.
- Saito, K., Wachi, A., Wataoka, K., & Akimoto, Y. (2023). Verbosity Bias in Preference Labeling by Large Language Models. arXiv:2310.10076. doi:10.48550/arXiv.2310.10076.
- Sciar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. arXiv:2310.11324. doi:10.48550/arXiv.2310.11324.
- Shen, W., Zheng, R., Zhan, W., Zhao, J., Dou, S., Gui, T., Zhang, Q., & Huang, X. (2023). Loose lips sink ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback. arXiv:2310.05199. doi:10.48550/arXiv.2310.05199.
- Singhal, P., Goyal, T., Xu, J., & Durrett, G. (2023). A Long Way to Go: Investigating Length Correlations in RLHF. arXiv:2310.03716. doi:10.48550/arXiv.2310.03716.
- Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. arXiv:2408.03314. doi:10.48550/arXiv.2408.03314.
- Song, S., Lee, J., & Ko, H. (2024). Hansel: Output Length Controlling Framework for Large Language Models. arXiv:2412.14033. doi:10.48550/arXiv.2412.14033.
- Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1). doi:10.2307/1412159.
- Sun, Y., Wang, H., Li, J., Liu, J., Li, X., Wen, H., et al. (2025). An Empirical Study of LLM Reasoning Ability Under Strict Output Length Constraint. arXiv:2504.14350. doi:10.48550/arXiv.2504.14350.
- Takase, S., & Okazaki, N. (2019). Positional Encoding to Control Output Sequence Length. arXiv:1904.07418. doi:10.48550/arXiv.1904.07418.
- Tam, Z. R., Wu, C.-K., Tsai, Y.-L., Lin, C.-Y., Lee, H., & Chen, Y.-N. (2024). Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models. arXiv:2408.02442. doi:10.48550/arXiv.2408.02442.
- Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., & Hupkes, D. (2024). Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. arXiv:2406.12624. doi:10.48550/arXiv.2406.12624.
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Large Language Models are not Fair Evaluators. arXiv:2305.17926. doi:10.48550/arXiv.2305.17926.
- Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., et al. (2024). Long-form factuality in large language models. arXiv:2403.18802. doi:10.48550/arXiv.2403.18802.
- Wu, Y., Wang, Y., Ye, Z., Du, T., Jegelka, S., & Wang, Y. (2025). When More is Less: Understanding Chain-of-Thought Length in LLMs. arXiv:2502.07266. doi:10.48550/arXiv.2502.07266.
- Xiao, M., Wang, A., Hu, Q., Miao, Z., Shen, H., Wang, L., Luo, W., & Su, J. (2026). Can LLMs Track Their Output Length? A Dynamic Feedback Mechanism for Precise Length Regulation. arXiv:2601.01768. doi:10.48550/arXiv.2601.01768.
- Xu, S., Xie, W., Zhao, L., & He, P. (2025a). Chain of Draft: Thinking Faster by Writing Less. arXiv:2502.18600. doi:10.48550/arXiv.2502.18600.
- Xu, Y., Dong, H., Wang, L., Sahoo, D., Li, J., & Xiong, C. (2025b). Scalable Chain of Thoughts via Elastic Reasoning. arXiv:2505.05315. doi:10.48550/arXiv.2505.05315.
- Yao, S., Chen, H., Hanjie, A. W., Yang, R., & Narasimhan, K. (2023). COLLIE: Systematic Construction of Constrained Text Generation Tasks. arXiv:2307.08689. doi:10.48550/arXiv.2307.08689.
- Yuan, W., Kulikov, I., Yu, P., Cho, K., Sukhbaatar, S., Weston, J., & Xu, J. (2024). Following Length Constraints in Instructions. arXiv:2406.17744. doi:10.48550/arXiv.2406.17744.
- Zhang, W., Zhou, Z., Wang, K., Fang, J., Zhang, Y., Wang, R., et al. (2025). LIFEbench: Evaluating Length Instruction Following in Large Language Models. arXiv:2505.16234. doi:10.48550/arXiv.2505.16234.
- Zhang, X., Cao, J., & You, C. (2024). Counting Ability of Large Language Models and Impact of Tokenization. arXiv:2410.19730. doi:10.48550/arXiv.2410.19730.
- Zhang, Y., Das, S. S., & Zhang, R. (2024). Verbosity != Veracity: Demystify Verbosity Compensation Behavior of Large Language Models. arXiv:2411.07858. doi:10.48550/arXiv.2411.07858.
- Zhao, H., Andriushchenko, M., Croce, F., & Flammarion, N. (2024). Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning. arXiv:2402.04833. doi:10.48550/arXiv.2402.04833.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685. doi:10.48550/arXiv.2306.05685.
- Zheng, M., Pei, J., Logeswaran, L., Lee, M., & Jurgens, D. (2024). When A Helpful Assistant Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models. arXiv:2311.10054. doi:10.48550/arXiv.2311.10054.
- Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023). Instruction-Following Evaluation for Large Language Models. arXiv:2311.07911. doi:10.48550/arXiv.2311.07911.
- Zhou, Q., Siddeeq, S., Saari, M., & Abrahamsson, P. (2026). A Closed-Loop Control Architecture for Reliable Constraint Satisfaction in LLM Text Generation. arXiv:2609.19710. doi:10.48550/arXiv.2609.19710.