PRIORBPB: REDUCING THE ANSWER LENGTH CONFOUND OF BITS-PER-BYTE WHILE IMPROVING ACCURACY PREDICTION

Anonymous authors
Paper under double-blind review

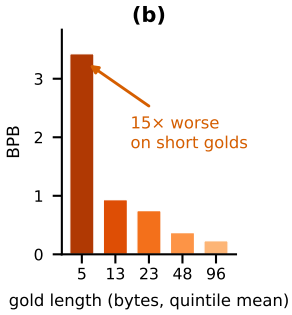
ABSTRACT

Bits-per-byte (BPB) normalises log-likelihood (in bits) by bytes rather than tokens to stay tokeniser-independent, and is widely used to compare language models across heterogeneous tasks. It also depends on how long the answer strings are. The bits a model spends on a gold string can be written as $\beta + \alpha/n + \varepsilon$, with α a fixed overhead, β a per-byte rate, (α, β) the least-squares fit in length, and ε the rest. Since the residuals sum to zero, total bits over total bytes is exactly $\beta + \alpha/\bar{n}$, with $\bar{n}$ the task's mean gold length. Because costs that do not scale with length are spread over the answer, tasks with long answers score better, as do long items within a task. Averaging per-item BPB has the same dependence. Across a 19-task suite on Qwen3 models, BPB correlates with mean gold length at Spearman -0.82 to -0.86. OLMo-3-7B behaves the same way (Spearman -0.79). Among the Qwen3 models we tested, the smallest model appears to score best (an apparent inverse scaling behaviour) due primarily to differences in the overhead. We propose a less length-biased metric, PRIORBPB, which estimates a per-(model,task) marginal per-byte cost curve from per-token surprisal and averages it under one shared length prior instead of under each task's own length distribution. This cuts the mean absolute Spearman correlation with answer length from 0.68 to 0.18 across 60 base checkpoints from five model families, while correlating with accuracy as well or better on base and instruct checkpoints (mean Spearman $0.50 \rightarrow 0.54$).

(a)

Prompt: "What is two plus two?"

response	Qwen3 4B	OLMo-3 7B
"4."	2.61	4.38
"It is four."	1.29	1.44
"The sum of two plus two is equal to 4."	0.63	0.88



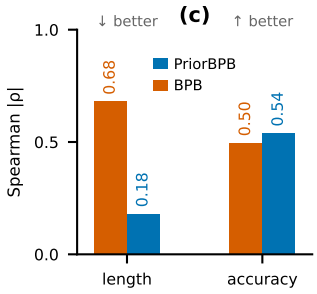


Figure 1: **(a)** Rephrasing the gold answer while keeping its semantics the same shifts BPB by 4–5× on Qwen3-4B and OLMo-3-7B-Instruct. **(b)** Within one task (MMLU full text answers) on Qwen3-4B, BPB is ∼15× worse on the shortest gold-length quintile than the longest. **(c)** PRIORBPB, which averages every model's per-byte cost under one shared length prior, is less length-confounded (mean $|\rho|$ with length $0.68 \rightarrow 0.18$, over 60 base checkpoints from five model families) and a better proxy for accuracy (mean ρ $0.50 \rightarrow 0.54$, those families' base and instruct checkpoints) than BPB.

1 INTRODUCTION

Leaderboards reduce a model to a few scalars, so a comparison is only as trustworthy as what those scalars measure and what comparisons they justify (Biderman et al., 2024; Liang et al., 2023;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Srivastava et al., 2023). For base models, a commonly-used cross-task scalar is *bits-per-byte* (BPB): the negative log-probability of a target answer or 'gold' continuation, summed over the answer and divided by its UTF-8 byte length (Gao et al., 2020; Magnusson et al., 2024; Gao et al., 2024). In bits-per-byte, next-token prediction is lossless compression and cross-entropy is the expected code length of the gold text (Delétang et al., 2024; Cover & Thomas, 2006; MacKay, 2003) (earlier work also reported per-character[1] bits (Radford et al., 2019)). BPB is (often) preferred over bits-per-token because the token denominator depends on the tokeniser and tokeniser fertility varies severalfold across scripts and languages (Petrov et al., 2023; Ahia et al., 2023), making bits-per-token incomparable across tokenisers, whereas bytes are a more 'fundamental' property of the text.

Changing this denominator, however, merely relocates the problem. Although bytes are tokeniser-independent, answer length is a property of the task and each of its samples, and any cost a model pays per item rather than per byte is spread over the length of each sample's gold continuation (which can also vary across samples within the task). If an item's bits are a fixed overhead α plus a rate β per byte, corpus BPB—total bits over total bytes—is exactly $\beta + \alpha/\bar{n}$, with $\bar{n}$ the task's mean gold length. This means that short- and long-answer tasks are scored differently, according to their differences in length—and also that short- and long-answer samples *within* tasks are scored differently. This is undesirable if BPB is to serve as a task- and sample-agnostic measure of predictive quality. Otherwise, it partly reflects the length composition of the benchmark and its samples rather than a difference in compression ability.

We investigate length bias in BPB, finding that answer length accounts for a large proportion of the score, both analytically and empirically (Section 3). This produces a clean-looking inverse-scaling result which is attributable to a fixed-cost effect (Section 4). We then introduce a drop-in less length-biased metric that uses every byte, is monotone, and reduces exactly to corpus BPB (Section 5). We validate it (Section 6) and condense the findings into a set of recommendations (Section 7).

Contributions. **(i)** A quantification of length bias effects in BPB. The identity $\beta + \alpha/\bar{n}$ shows the score can be largely read as a per-byte rate plus a fixed cost over mean length. We quantify how strongly the fixed-cost term accounts for cross-task and within-task comparisons. **(ii)** A case study of apparent inverse scaling behaviour that is an artefact of the fixed-cost term. **(iii)** A second, compounding error introduced by the mean-of-ratios metric, which inflates BPB by up to 99% compared to corpus BPB. **(iv)** PRIORBPB, a less length-biased metric built on a per-(model,task) cost curve and a shared length prior, with exact reductions and checks of its length behaviour across various models and tasks. **(v)** Evidence that, across dense Pythia pretraining checkpoints and a cross-tokeniser and -model sweep, PRIORBPB tracks task accuracy at least as well as mean-of-ratios BPB on base checkpoints and better on instruct checkpoints, while decreasing the length confound (Appendix K). **(vi)** A set of recommendations for when and how to use BPB variants.

2 BACKGROUND AND SETUP

Bits-per-byte. For a prompt x and gold continuation c of byte length n, let $L = -\log_2 P_\theta(c \mid x)$ be the total surprisal in bits under a model with parameters θ. A task has N items indexed by i, giving pairs $\{(L_i, n_i)\}_{i=1}^N$. The *corpus* BPB form $\sum_i L_i / \sum_i n_i$ is a compression rate and can also be considered a classical ratio estimator $\sum_i Y_i / \sum_i X_i$ from survey sampling (Cochran, 1977) with $Y_i = L_i$ and $X_i = n_i$. The *mean-of-ratios* BPB form $\frac{1}{N}\sum_i L_i/n_i$ is what popular evaluation frameworks emit as "`BitsPerByte`" (Gao et al., 2024); it is the mean of per-item rates, which is not itself a rate. It overweights short items and inflates the reported rate. Anticipating the affine model of the next paragraph, write $L_i = \alpha + \beta n_i + \varepsilon_i$, where α and β are fitted once per (model, task) across that task's N items, β is the per-byte rate, and ε_i absorbs whatever else about item i that is unrelated to its length, e.g., intrinsic difficulty. Then

$$\frac{1}{N}\sum_i \frac{L_i}{n_i} = \underbrace{\beta + \frac{\alpha}{\bar{n}}}_{\text{corpus BPB}} + \underbrace{\alpha\left(\frac{1}{N}\sum_i \frac{1}{n_i} - \frac{1}{\bar{n}}\right)}_{\geq 0 \text{ for } \alpha \geq 0} + \underbrace{\widehat{\mathrm{Cov}}(\varepsilon_i,\, 1/n_i)}_{\text{unsigned}}.$$

[1] Characters are a less well-defined unit than bytes, since the count depends on whether one means code points or grapheme clusters. Unicode normalisation can also change the count for the same rendered text, whereas the UTF-8 byte length of a string is fixed by the encoding.

The second term is non-negative by Jensen's inequality for the convex map $n \mapsto 1/n$, and is zero only when every item has the same length. The third term is a covariance because the least-squares residuals have zero mean; it vanishes in corpus BPB, where the residuals enter with equal weight and sum to zero, but not in mean-of-ratios BPB, which weights each residual by $1/n_i$. Mean-of-ratios therefore inherits the fixed-cost confound and adds two terms of its own.

An affine model of cost. Oostermeijer (2026) observes, across tokenisers and benchmarks, that summed log-likelihood is approximately affine in answer length, i.e. that ε_i is small relative to $\alpha + \beta n_i$. We interpret α as a length-independent fixed cost, dominated by the first answer token, which starts a span with weak local context, and β as the asymptotic per-byte continuation rate. We use (α, β) to denote the ordinary-least-squares fit. We emphasise that all claims which make any reference to α or β should be read with the caveat "under the least-squares affine fit to length" in mind. We use the decomposition only to isolate the length confound and make no attempt to model ε in the current work. Because cost curves have curvature near the first bytes, when we later introduce PRIORBPBin Section 5, we use a non-parametric equation to correct for the length bias rather than relying on the two-parameter fit of (α, β).

Related length corrections. Byte- or token-length normalisation (Brown et al., 2020) divides by length, in what turns out empirically to be an ironic attempt to minimise length bias. Oostermeijer (2026) shows such normalisation frequently over-corrects, and proposes *Bayesian accuracy*, which imposes an explicit exponential prior over answer length and subtracts a fitted length trend ($\ell_\theta(c \mid x) - bn$) from candidate answers before selecting the most 'likely' answer as the model's answer. Pointwise mutual information (Fano, 1961; Holtzman et al., 2021) and its asymmetric normalisation (Cho et al., 2026) subtract an unconditional log-likelihood but require extra forward passes, and answer-prior calibrations correct format and surface-form biases (Zhao et al., 2021). All of these rank candidate answers *within* a task. We instead want a per-byte quality scalar *comparable across tasks*, which (as we show) calls for a global rather than per-task prior. The closest, Bayesian accuracy (Oostermeijer, 2026), subtracts a fitted length trend to choose the answer within a task. PRIORBPB instead keeps the bits-per-byte scale as a cross-task quality scalar, shares a single prior across tasks (a per-task prior recovers the confounded corpus number exactly, Section 5), and reduces to corpus BPB under the self-prior, the reduction behind the structural results of Appendix B.[2]

Experimental setup. We score two model families on the same 22 tasks: Qwen3 (Yang et al., 2025) at 0.6B, 1.7B, and 4B, and OLMo-3 (Team OLMo, 2026) at 7B. The evaluation suite includes letter-label multiple choice (ARC/MMLU under the OLMES protocol (Gu et al., 2024; Clark et al., 2018; Hendrycks et al., 2021a)), extractive QA (SQuAD2 (Rajpurkar et al., 2018), NaturalQuestions (Kwiatkowski et al., 2019)), sentence completion (HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2019), WinoGrande (Sakaguchi et al., 2019)), science and knowledge (SciQ (Welbl et al., 2017), MMLU), mathematics (GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b)) and code (MBPP (Austin et al., 2021), HumanEval (Chen et al., 2021)), along with seven German counterparts. Golds are full-text continuations except the two letter-label multiple choice tasks. We randomly select 500 items per task, except HumanEval (164) and MATH (3,453 via per-subject sampling). We compute, per item, $(L_i, n_i, \text{tokens})$, per-token surprisal, and per-token byte lengths.

Statistical inference. We report Bayesian summaries throughout (credible intervals and posterior probabilities). Our estimator likewise aggregates against an explicit length prior, though it is a prior-weighted average rather than a posterior. The full inference protocol (Bayesian bootstrap posteriors, Savage–Dickey Bayes factors reported at both the task and task-family counts, clustered resampling, and a prior-width sweep) is provided in Appendix H.

[2]Although we do not address it in the current work, we note that the same confound appears in generation evaluation, where longer outputs are systematically preferred and must be debiased (Singhal et al., 2023; Park et al., 2024; Dubois et al., 2024), which may be a confound further amplified in preference-based post-training (Rafailov et al., 2023; Ouyang et al., 2022).

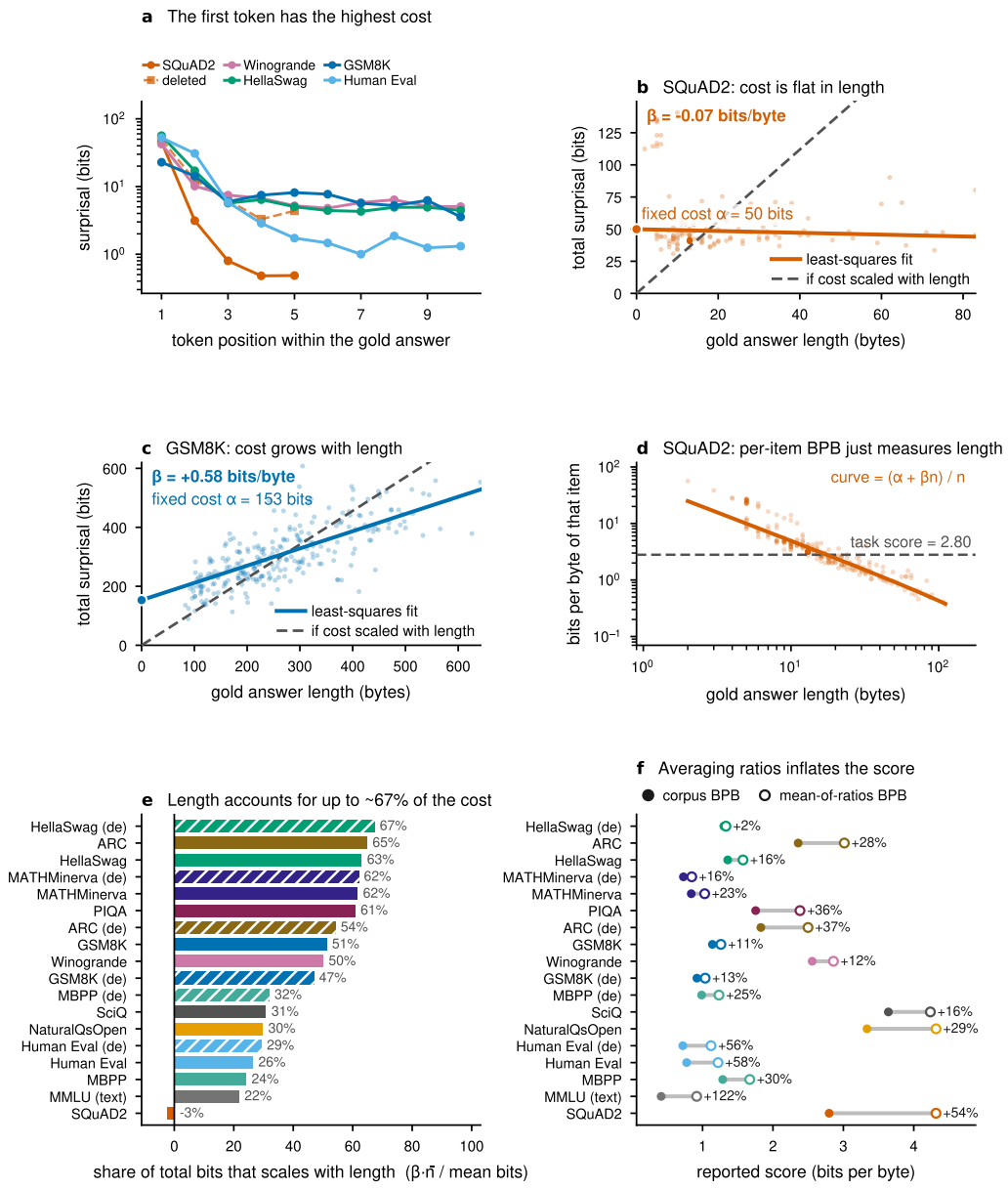


Figure 2: The length confound on Qwen3-4B over 19 evaluations. **(a)** Mean surprisal by token position (log scale): the first gold token costs an order of magnitude more than later tokens; for SQuAD the later tokens are apparently nearly 'free' because they are copied from the passage—whereas, when the answer-bearing sentence (dashed) is deleted, the later tokens become more expensive. **(b,c)** Total surprisal vs. answer length with the least-squares fit. The intercept α is a fixed cost accounting for almost the entire score in (b) and about half of it in (c). **(d)** Because the cost barely grows with length, dividing it by n leaves roughly α/n, i.e., the shortest answers score $> 30\times$ higher than the longest. **(e)** Length accounts for up to approximately 67% of the cost, depending on the task, so no single normalisation fits all tasks. **(f)** Mean-of-ratios BPB weighs short items by $1/n$ and inflates the score by up to 122%. Appendix D shows qualitatively-similar results for Qwen3-0.6B and OLMo-3-7B-Instruct.

3 BITS-PER-BYTE IS A LENGTH STATISTIC

An exact identity. Because the least-squares residuals of $L_i \approx \alpha + \beta n_i$ sum to zero, corpus BPB satisfies

$$\frac{\sum_i L_i}{\sum_i n_i} \;=\; \frac{\alpha + \beta\,\bar{n}}{\bar{n}} \;=\; \beta + \frac{\alpha}{\bar{n}}, \tag{1}$$

where $\bar{n}$ is the task's mean gold length. The identity is algebraic, so it holds whatever the residuals and we verify Eq. equation 1 holds to 2×10^{-15} across all 57 model×task fits. Any fixed cost α contributes a term $\frac{\alpha}{\bar{n}}$ to BPB that decays hyperbolically with the task's mean gold length. So, two tasks with identical per-byte difficulty β but different $\bar{n}$ receive different BPB purely through $\alpha/\bar{n}$.

Cross-task BPB tracks length rather than difficulty. Across the 19 length-varying tasks, corpus BPB correlates with mean gold length at Spearman -0.77 (95% credible interval $[-0.92, -0.49]$, posterior $P(\rho < 0) > 0.999$). As a prior-model sensitivity check, we calculate the Fisher-z Bayes factor, which concurs: $\mathrm{BF}_{10} \approx 660$ (extreme), robust to the prior ($\mathrm{BF}_{10} \geq 100$ for all $\sigma_0 \in [0.3, 2]$), and still strong once the 19 tasks are grouped into their 12 capability domains ($\mathrm{BF}_{10} \approx 22$ at $n_{\mathrm{eff}} = 12$). By comparison, mean-of-ratios BPB correlates even more strongly with mean gold length, at -0.83. Decomposing the least-squares fit indicates the fixed-cost part $\alpha/\bar{n}$ correlates -0.71 with length ($\mathrm{BF}_{10} \approx 84$, strong evidence) while for the length-free rate β (-0.34) the evidence is inconclusive and if anything leans null ($\mathrm{BF}_{01} \approx 1.6$) (Fig. 7, left). These results are not Qwen3-specific; on OLMo-3-7B-Instruct, scored on the same 19 length-varying tasks, corpus BPB correlates with mean gold length at Spearman -0.67 (95% credible interval $[-0.87, -0.37]$; mean-of-ratios BPB -0.79), with the correlation again higher for $\alpha/\bar{n}$ (-0.58) than β (-0.31). Additionally, short items score worse than long items on all tested tasks (median $2.1\times$; Fig. 7, right). On Qwen3-0.6B, the mean-of-ratios BPB for MMLU-text being 0.22 compared to SQuAD2 being 2.39 is therefore mostly a statement about the distribution of their gold lengths, and says little about their relative difficulty. Both scores are also individually biased within their tasks by the diversity of samples length distributions.

How robust is the effect? A sceptic might argue long-gold tasks are genuinely more predictable. To check this, we hold task and model fixed and bin a task's own items by gold length. Comparing the shortest fifth to the longest fifth, the score is worse on the short items on 19 of 19 tasks, by a median of $2.3\times$. This comports with the per-token traces, which show the first gold token accounting for 7–50 bits and later tokens far less (Fig. 2a, Appendix Fig. 9). On SQuAD the later tokens cost almost nothing, and deleting the answer-bearing sentence from the prompt raises the total by $\sim$15 bits, so its flat $\beta \approx 0$ is possibly a 'copyability' effect rather than those bytes being easy to produce. At matched answer byte positions, the per-byte cost of long and short golds is indistinguishable (median ratio 0.98, 95% credible interval $[0.94, 1.01]$; Section 6). BPB therefore conflates the construct we intend to measure (per-byte difficulty) with a nuisance confound (total answer length). Since different evaluation tasks have different length distributions (Fig. 2e), no single global length normalisation can accommodate all simultaneously.

4 A CASE STUDY: INVERSE SCALING THAT ISN'T

Here we show how the answer length confound can produce a misleading conclusion w.r.t. capability trends across models. Inverse scaling, where larger models are reliably worse, is an interesting and sometimes surprising phenomenon, with several identified mechanisms (McKenzie et al., 2023; Wei et al., 2023). However, such a trend could also be an artefact of the chosen metric rather than of capability (Schaeffer et al., 2023; Wei et al., 2022), and here we show this latter kind. Under corpus BPB the smallest model, Qwen3-0.6B, attains the lowest (best) score on 15 of 19 tasks, an apparent inverse-scaling law. However, upon decomposing the cost, we find that the fixed cost α is larger at 4B than at 0.6B on all 19 tasks (e.g., GSM8K $22 \rightarrow 93 \rightarrow 153$ bits), while the per-byte rate β shows no such trend (Appendix Fig. 6). Appendix E gives the full experiment and per-task details. We hypothesise that these differences in the fixed cost arise from differences in the task prompts and their formats, which are known to have first-order effects on measured performance (Sclar et al., 2024). Since these completion-style golds are wrapped in an instruct chat template, a larger, better instruction-tuned model might pay more to emit the bare continuations required in

these benchmarks. A controlled replicate (`enable_thinking=False`, following the Qwen3 chat template (Yang et al., 2025)) still shows this anomaly (0.6B has the highest performance on 18 of 19 tasks). Independently, the mean-of-ratios BPB can have a larger fixed-cost term than corpus BPB (since $\mathbb{E}[1/n] \geq 1/\bar{n}$) and can dramatically inflate scores. On this basis, we later recommend generally preferring corpus BPB over mean-of-ratios BPB.

5 A LESS LENGTH-BIASED ESTIMATOR THAT USES EVERY BYTE

Design goal. We want to measure per-byte continuation difficulty: the bits the model needs per gold byte, whatever the answer's length. The rate β is the obvious candidate, since Eq. equation 1 puts the entire length confound in the other term. It is not usable as a score. It is the slope of a straight line through a cost curve that bends sharply over its first few bytes (Fig. 8), so it describes the tail of an answer and discards whatever the fit leaves in ε; where a model pays most of its bits up front, it even goes negative ($\beta = -0.07$ bits/byte on SQuAD2 at 4B). We therefore want a per-byte quality scalar that (i) is comparable across tasks with different length distributions, (ii) uses every byte of every answer, (iii) is monotone, so that making any answer more probable cannot worsen the score, and (iv) reduces to corpus BPB under a stated condition, so its behaviour is interpretable.

From per-token surprisal to a cost curve. The object we estimate is the average cost of the k-th byte of an answer: the bits a model spends there, averaged over the items long enough to have a k-th byte. Because tokens and bytes do not align, we first spread each token's bits linearly across the bytes it spans, giving a cost $c_i(k)$ for each of the n_i answer bytes of item i. The *marginal cost curve* averages these by position, and $A(k)$ accumulates them:

$$R(k) \;=\; \operatorname*{mean}_{i:\, n_i \geq k} c_i(k), \qquad A(k) \;=\; \sum_{j \leq k} R(j). \tag{2}$$

One marginal cost curve is estimated per (model, task), and every gold answer contributes to every position it reaches, i.e., byte k contributes to $R(k)$, e.g., a 3-byte gold answer contributes to $R(1)$, $R(2)$, and $R(3)$. Since no item's gold answer is dropped or truncated, and since $R(k)$ averages non-negative costs with non-negative weights, making any byte more surprising can only raise the score (Prop. 6). This therefore achieves design goals (ii) and (iii) above. The only limitations is that at large k the average runs only over items that reach byte k, so late positions describe a length-selected subpopulation. We discuss this potential survivorship bias in Appendix A.

Aggregation under a shared length prior. The problem we have now is that $R(k)$ is a curve but we want a scalar score. Choosing how much weight each byte position gets in this score is therefore our goal, e.g., corpus BPB lets position k count once for every item that reaches position k, so each task weights positions by its length distribution (which is the confound we are trying to avoid). We therefore set and apply the same weights for all positions, and following the 'explicit prior, subtract not divide' logic of Bayesian accuracy (Oostermeijer, 2026), we weight position k by $s(k) = P_{\text{prior}}(N \geq k)$, the probability that an answer drawn from a fixed *length prior* runs to at least k bytes, and report the weighted average

$$\textsc{PriorBPB} \;=\; \frac{\sum_k s(k)\, R(k)}{\sum_k s(k)} \;=\; \frac{\mathbb{E}_{\text{prior}}[A(N)]}{\mathbb{E}_{\text{prior}}[N]}. \tag{3}$$

We use a geometric prior, $s(k) = (1 - 1/\mu)^{k-1}$, with mean length μ frozen at 111 bytes, the pooled median answer length across the suite. Every position keeps positive weight, so no gold answer's late bytes drop out of the score; they are weighted by the prior rather than by how many items in their own task happen to be that long. PriorBPB therefore reports every model's per-byte cost at the same declared answer-length profile, which is what makes the numbers comparable across tasks. Since μ comes from the length distribution alone, fixed before any score is computed, it cannot be tuned to the length or accuracy results, and it transfers to held-out tasks and further model families (Appendix J).

Key reduction. Setting $s(\cdot)$ to the task's own empirical length distribution gives $\sum_k P(N \geq k)R(k) = \frac{1}{N}\sum_i L_i$ and $\mathbb{E}[N] = \bar{n}$, so PriorBPB = corpus BPB exactly; debiasing only swaps this self-prior for a shared one, and because the byte denominator is kept, PriorBPB inherits

the tokeniser-independence that motivates BPB. Corpus BPB and PRIORBPB are thus the same weighted average of one cost curve, differing only in whether the position weight is read off each task's own lengths or fixed in advance; two elementary structural facts then follow (Appendix B): no non-degenerate score is length-invariant, and a model comparison is prior-independent exactly when one cost curve pointwise dominates the other, so the scale 'anomaly' of Section 4 reflects crossing cost curves rather than a robust capability gap. Appendix B also gives the geometric closed form, the domain/task reading, the shared-vs-per-task-prior control, and the PREFIXBPB diagnostic slice.

6 PROPERTIES OF LESS LENGTH-BIASED ESTIMATION

We fix the tunable choices (K, μ, and the prior width σ_0) before computing any outcome, then check the less length-biased estimators against corpus BPB on the complete per-token re-score of all three Qwen3 scales (0.6B/1.7B/4B, all 22 tasks; Fig. 3). We report the 0.6B numbers with the 1.7B/4B replication alongside, and every result holds on all three scales. The accuracy comparison below is held to length-corrected and length-independent control answers, so a length artefact cannot be mistaken for an advantage. The two main properties are within-task length invariance and cross-task deconfounding.

Within-task length invariance. The top/bottom length-quintile ratio moves from 0.43 (corpus, strongly biased towards long golds) to 1.10 (PREFIXBPB), a slight residual tilt in the opposite direction, but $\sim 5\times$ smaller than corpus's deviation and shrinking with scale to $1.08/1.04$ on 1.7B/4B. The 95% credible interval 0.99–1.33 still covers the unbiased value 1, though it sits almost entirely above it (posterior probability 0.50 of lying within $\pm 10\%$ of 1), so we read this as a mild, decaying over-correction rather than exact invariance (Prop. 5). PRIORBPB applies the same test at the per-item level (comparing the shared-support rate curve $R(k)$ of the long- versus short-gold quintiles under the fixed prior) and lands even closer to neutral (median 0.98, 95% credible interval $[0.94, 1.01]$), confirming that at a fixed content position the per-byte cost is essentially independent of a gold's eventual length (Fig. 3, left).

Cross-task deconfounding. The length correlation falls from -0.78 (corpus; posterior $P(\rho < 0) > 0.999$) to -0.34 (PRIORBPB, shared prior; 95% credible interval $[-0.67, +0.04]$); a paired contrast places the shared prior below corpus in length dependence with posterior probability > 0.999 ($\Delta|\rho| = 0.41$), while the per-task-prior control is back at -0.74. The deconfounding replicates on 1.7B and 4B (shared-prior $\rho = -0.14$ and -0.31; paired $\Delta|\rho| = 0.53$ and 0.44, $P \geq 0.996$; Fig. 3, right).

Further properties and generalisation. The following properties hold by construction or by wide margins empirically for PRIORBPB. The estimator is monotone, i.e., making any token more surprising never lowers the score. This is guaranteed by non-negative weights (see Prop. 6). It is statistically efficient, with estimator noise comparable to corpus BPB (CV ratio 0.97). It is cross-lingually neutral: because token counts for the same content differ markedly across languages (Petrov et al., 2023; Ahia et al., 2023), on matched German/English pairs the median $|\text{DE/EN} - 1|$ improves from 0.100 to 0.027, and on Qwen3 bits-per-byte is closer to language-neutral than bits-per-token on 8/12 gold-translated measurements (0.14 vs. 0.21; Fig. 10). This byte-over-token advantage is tokeniser-specific, however, e.g., on a second tokeniser (SmolLM2 (Ben Allal et al., 2025)) the byte denominator is actually the less language-stable of the two (Appendix I). Linear vs. token-boundary byte attribution also agree to 10% and the leading-space convention moves the score by 4% (Prop. 11). We also find that these length results generalise on further model families and transfer to held-out tasks with the prior, μ, kept fixed (Appendix J).

PRIORBPB is both less length-biased and a better accuracy proxy. PRIORBPB is better in two senses. As a measurement, PRIORBPB is less length-confounded than mean-of-ratios BPB (the length properties above), so it is the more honest per-byte scalar. As a proxy, it also tracks downstream accuracy more faithfully. In a paired, clustered-Bayesian comparison against the mean-of-ratios aggregator across dense Pythia pretraining ladders (clean sizes $\times$ log-spaced steps) and a cross-tokeniser final sweep (Appendix K, Table 1), the paired fidelity advantage $\Delta = \text{fidelity}(\text{PRIORBPB}) - \text{fidelity}(\text{mean-of-ratios})$ (higher = closer tracking of accuracy along a ladder) is reliably positive: $\Delta = +0.021$ on base checkpoints ($P(\Delta > 0) = 1.000$, 95% CrI

Table 1: PRIORBPB tracks downstream accuracy better than the mean-of-ratios BPB reported by evaluation frameworks (Gao et al., 2024). Paired superiority endpoint Δ = fidelity(PRIORBPB) − fidelity(mean-of-ratios); primary gold raw accuracy, statistic −Spearman; clustered Bayesian bootstrap. We report the posterior mean, 95% CrI; a comparison is reliably signed when the interval excludes zero. Full set of models, golds, and per-gold robustness checks are in Appendix K.

Models	n	Δ (posterior mean)	95% CrI
Base	62	$+0.021$	$[+0.004, +0.079]$
Instruct	19	$+0.122$	$[+0.005, +0.571]$
Pooled	81	$+0.047$	$[+0.011, +0.163]$

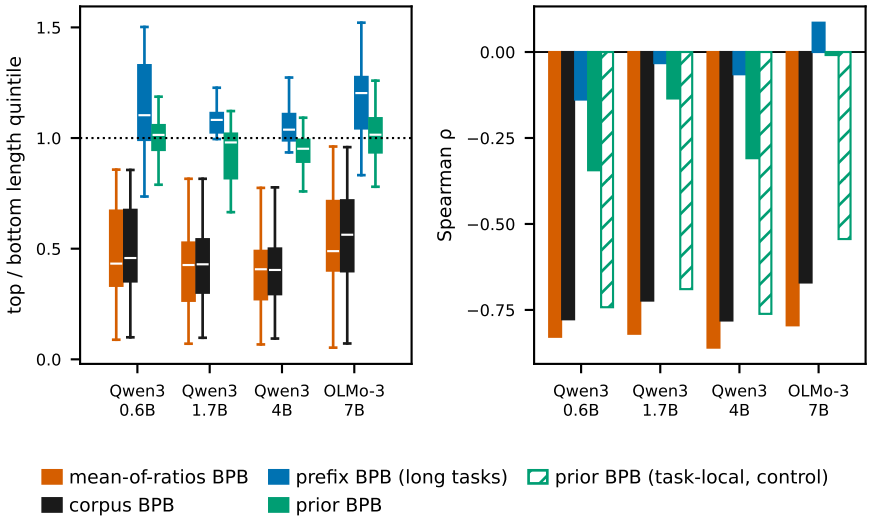


Figure 3: Length behaviour of the estimators on Qwen3-instruct (0.6B, 1.7B, 4B) and OLMo-3-7B-instruct. **Left, within-task length invariance:** top/bottom gold-length quintile ratio (1 is unbiased). Both estimators are length-invariant: PREFIXBPB through its first-K-byte instrument and PRIORBPB through the shared-support rate curve $R(k)$ (median 0.98, 95% credible interval $[0.94, 1.01]$), whereas corpus BPB and mean-of-ratios BPB tilt to ≈ 0.43 (long golds under-costed). **Right, cross-task deconfounding:** Spearman(score, mean gold length) across every non-label task (0 is deconfounded). PRIORBPB with the shared prior roughly halves the corpus correlation while using every byte; the task-local-prior control (hatched) collapses back onto corpus. PREFIXBPB is undefined on short-gold tasks, so its right-panel bar (blue) is computed on the long-task subset only (16 of 19 tasks) and is not directly comparable to the all-task bars: there its near-zero correlation simply reflects that truncating every gold to a fixed byte budget discards essentially all remaining length variation, which is also why PRIORBPB (which instead uses every byte) is the more informative cross-task estimator. The cross-lingual neutrality check, including its per-estimator German/English ratios, is in Fig. 10. Each panel groups the same four models.

$[+0.004, +0.079]$), a larger $\Delta = +0.122$ under instruct/templated scoring ($P(\Delta > 0) = 1.000$, interval excluding zero on all three fidelity statistics), and $+0.047$ pooled. The instruct gain is the larger of the two because templated scoring injects exactly the length-varying fixed cost PRIORBPB removes, so debiasing helps most in the regime deployed evaluation uses; on base checkpoints, where that fixed cost is smaller, the advantage is reliably signed but modest ($+0.021$) and we do not lean on it. For scale, the fidelity gap between the two incumbent estimators (mean-of-ratios BPB vs. corpus BPB) is ~ 0 (base) and -0.031 (instruct), so PRIORBPB's advantage already exceeds the entire spread among the estimators in use today, more so on instruct. It is never worse on any stratum or gold and ties exactly on length-independent letter-label accuracy (where the two estimators coincide by construction), so the gain survives length correction and is not a shared-length artefact. The full set of models, per-gold robustness checks, and per-stratum summary table are in Appendix K.

7 RECOMMENDATIONS FOR USING BPB VARIANTS

Based on our results, we make the following recommendations. **(i)** Do not naïvely average mean-of-ratios or corpus BPB across tasks, as it weights tasks by length. **(ii)** Report corpus BPB only within a fixed task when you are happy to assume a shared fixed cost α across models, since the length mix will also be shared across the models and $\alpha/\bar{n}$ will be a common offset. **(iii)** For a cross-task scalar comparison, use PRIORBPBwith a frozen, shared length prior, and state the prior. **(iv)** Never report mean-of-ratios BPBwithout a (α, β) decomposition alongside is to differentiate fixed costs from per-byte costs and, as Section 4 shows, expose format-mismatch and scaling artefacts, e.g., do not read a series of mean-of-ratios BPB results as a scaling law (Kaplan et al., 2020; Hoffmann et al., 2022; McKenzie et al., 2023) without checking α or β. Mean-of-ratios BPB is not a rate and can inflate scores by $\sim 2\times$.

When debiasing matters. Raw corpus BPB is acceptable to use as a within-task statistic and needs no debiasing. Technical reports that quote a cross-task BPB should publish both the raw corpus number and PRIORBPB under a stated shared prior. PRIORBPB is worth its added complexity only when a single cross-task scalar is unavoidable, since the (α, β) decomposition can only serve as a diagnostic indicator of length bias.

When does debiasing matter? A possible objection is that raw BPB numbers computed on the same dataset are already comparable, so why attempt to debias them? Within a single task the length mix is a property of the data shared by every model, so $\alpha/\bar{n}$ is a common additive offset and model rankings on that task are unaffected by the confound. In that case, so long as you are willing to ignore potential differences in the fixed cost α, raw corpus BPB can be a suitable within-task statistic. The length confound is much more problematic as soon as you want to evaluate a number number of different tasks, since: (i) in any cross-task average or leaderboard, where per-task offsets differ, standard BPB weights tasks by length; (ii) you are implicitly reading a per-byte number as a capability rather than a within-task ranking, since the fixed cost is not continuation difficulty; and (iii) you are comparing different-length variants of one capability (the same task in a second language, or a format that lengthens answer strings), where the length distribution itself changes. A controlled within-task length stratification (Appendix L) isolates (iii) by holding the task fixed and varying only the gold answer length distribution across strata. Nevertheless, PRIORBPB is reliably the more length-neutral estimator.

A training-objective interpretation. BPB is cross-entropy measured per byte, and cross-entropy is what pre-training minimises, so the length confound is in the evaluation distribution we aggregate over rather than the objective, which is why we propose to fix the aggregation method rather the loss. The same separation recurs in post-training, where reward models and preference optimisation (Ouyang et al., 2022; Rafailov et al., 2023) are documented to entangle quality with response length, to the point that a purely length-based reward reproduces most RLHF gains (Singhal et al., 2023; Park et al., 2024), and Kalai et al. (2025) argue that cross-entropy pre-training yields a calibrated base model while post-training optimised against mis-specified graders drifts from it. Technical reports and leaderboards that quote a single cross-task BPB should therefore ideally seek to publish both the raw corpus number and PRIORBPB under a stated shared prior rather than raw cross-task averages.

8 CONCLUSION

Bits-per-byte is commonly used as a text-intrinsic, tokeniser-independent measure of model quality, but is heavily confounded by gold lengths, which makes comparisons between tasks largely comparisons of length distributions. In one case, this confound is severe enough to imply a false inverse-scaling capability trend. Our proposed replacement, PRIORBPB, states an explicit, shared length prior and integrates a per-byte cost curve against it, rather than dividing by length. This recovers corpus BPB as the (biased) self-prior special case and reduces the confound while using every byte. We note these conclusions come with limitations and residual caveats (which we outline in Appendix A). What a metric measures, and what inference it supports, should be stated and tested; we hope the identity, the case study, and our recommendations are helpful for increasing trust in language model evaluation.

AI USE STATEMENT

In this work, we used generative AI tools to assist us in implementing our methods in Python code, i.e., agentic coding assistants. Code was checked for correctness and code quality by multiple human authors, and was also verified and tested for correctness programmatically. We have not used generative AI tools for other tasks with required disclosure. Additionally, we used generative AI tools to assist us in finding relevant literature and giving feedback on writing to improve readability. We have reviewed all AI-assisted work, e.g., we manually edited and adapted LLM-suggested wording into our text. We take responsibility for the final content of this work, including text, claims, or artefacts produced with the aid of generative AI.

ETHICS STATEMENT

We collected no human-subject or personally identifiable data in this work. This is a methodological paper and raises no ethical concerns beyond those already present in the benchmarks it re-analyses. It identifies a length confound in bits-per-byte that affects a wide array of evaluations. All datasets used (ARC, MMLU, SQuAD2, GSM8K, MATH, MBPP, HumanEval, and their German counterparts, among others) are public research benchmarks. The models evaluated (Qwen3, OLMo-3, Pythia, Qwen2.5, OLMo-2, SmolLM2) are open-weight models used under their existing licenses only for scoring/evaluation. We believe aiming to reduce the length bias of evaluations can make published model comparisons harder to game and more trustworthy.

REPRODUCIBILITY STATEMENT

We provide an anonymised copy of the evaluation code as a supplementary zip archive. The main text defines corpus, prefix, and prior bits-per-byte and states how they are aggregated. The code contains the corresponding task and metric implementations, together with REVIEWING_README.MD, which shows how to run MMLU and read the three metric results from the generated results aggregate file. All other results for different tasks and models can be reproduced in the same way. The benchmarks used in that check are public datasets loaded by those task implementations. All assumptions and set-up/notation is provided in the main text and appendix for parsing the theoretical proofs.

AUTHOR CONTRIBUTIONS

Redacted for anonymity.

ACKNOWLEDGMENTS

Redacted for anonymity.

REFERENCES

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on empirical methods in natural language processing*, pp. 9904–9923, 2023.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79, 2010.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Pavel Chizhov, Mattia Nee, Pierre-Carl Langlais, and Ivan P. Yamshchikov. What the HellaSwag? on the validity of common-sense reasoning benchmarks. *arXiv preprint arXiv:2504.07825*, 2025.

Gyeongje Cho, Yeonkyoung So, and Jaejin Lee. Choices speak louder than questions, 2026. URL https://arxiv.org/abs/2502.18798.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, 3rd edition, 1977.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *International Conference on Learning Representations (ICLR)*, 2024.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Robert M. Fano. *Transmission of Information: A Statistical Theory of Communications*. M.I.T. Press, 1961.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. *arXiv preprint arXiv:2406.08446*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021b.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. *arXiv preprint arXiv:2104.08315*, 2021.

Harold Jeffreys. *Theory of Probability*. Oxford University Press, 3rd edition, 1961.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

David H. Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky. *Foundations of Measurement, Volume I: Additive and Polynomial Representations*. Academic Press, 1971.

John K. Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 2018.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics (TACL)*, 7, 2019.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check. *arXiv preprint arXiv:2507.00885*, 2025.

Alexander Ly, Josine Verhagen, and Eric-Jan Wagenmakers. Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 2016.

David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. Paloma: A benchmark for evaluating language model fit. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, et al. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research (TMLR)*, 2023.

Koen Oostermeijer. Accuracy and normalized accuracy under length bias: Analysis, guidelines, and a bayesian alternative. In *International Conference on Machine Learning (ICML)*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 2010.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*, 2018.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Donald B. Rubin. The bayesian bootstrap. *The Annals of Statistics*, 9(1), 1981.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design, or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. *arXiv preprint arXiv:2310.03716*, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2019.

Team OLMo. OLMo 3. *arXiv preprint arXiv:2512.13961*, 2026.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022.

Jason Wei, Najoung Kim, Yi Tay, and Quoc Le. Inverse scaling can become U-shaped. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15580–15591, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.963. URL https://aclanthology.org/2023.emnlp-main.963/.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*, 2017.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Guanhua Zhang, Ricardo Dominguez-Olmedo, and Moritz Hardt. Train-before-test harmonizes language model rankings. In *International Conference on Learning Representations (ICLR)*, 2026.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning (ICML)*, 2021.

A Limitations

Our case-study magnitudes use Qwen3-instruct at three scales, so absolute values are specific; the confound, however, is an algebraic identity (Eq. equation 1) and therefore family-independent. The problem and the within-task and cross-task length checks replicate on OLMo-3-7B-Instruct (same 22-tasks), on the complete per-token re-score of all three Qwen3 scales, on two further families (SmolLM2-base and Qwen3-Base), and on a held-out suite of datasets under a frozen prior (Appendix J). The affine fits are an approximation (real cost curves have curvature near the first bytes), which is why we use the non-parametric $R(k)$ of Eq. equation 2 rather than the two-parameter fit. There are two further caveats to $R(k)$ and to the byte denominator. First, $R(k)$ at position k averages only over items that reach byte k, so its late-position values describe a length-selected subpopulation; this survivorship is one plausible source of the residual length correlation (e.g., -0.34 at 0.6B), and the fixed prior mitigates but does not completely remove it. An isolation check (Appendix H) locates the residual in the deep tail rather than in short-item survivorship: dropping short items entirely leaves the residual at -0.34, whereas capping the integration at $k \leq 64$ bytes (which discards exactly the sparsely-supported tail positions) drives it to ≈ 0 (paired advantage $+0.62$). Second, the byte denominator's cross-lingual advantage is tokeniser-specific: on a second tokeniser (SmolLM2)

bits-per-token is the more language-stable scale (Appendix I), so byte-normalisation should not be assumed language-neutral for every tokeniser. The format-mismatch explanation for the intercept's growth with scale (Section 4) is supported (the base-checkpoint control of Appendix I attenuates the fixed-cost growth from 100% to 25% of tasks once the chat template is removed) but not proven, as base and instruct checkpoints differ in more than the template.

B A LESS BIASED ESTIMATOR: REDUCTIONS AND STRUCTURE

Closed-form reductions. The self-prior reduction of Section 5, which recovers corpus BPB when the prior is each task's own length distribution, has a companion closed form. Flat cost gives that constant for any prior. For the geometric prior of mean μ and affine cost $C(k) = \alpha + \beta k$, the exact value over a position budget of the longest gold $n_{\max}$ is $\textsc{PriorBPB} = \beta + \alpha/\big(\mu\,(1-(1-1/\mu)^{n_{\max}})\big)$, which decreases to the untruncated $\beta + \alpha/\mu$ as golds lengthen: truncation inflates the fixed cost by $1/(1-(1-1/\mu)^{n_{\max}}) \geq 1$. Either way the task-varying $\alpha/\bar{n}$ confound of Eq. equation 1 becomes a shared α/μ (up to this mild, length-dependent truncation factor, which is itself part of the residual dependence measured below) while the all-byte rate β is preserved.

A domain/task reading. The self-prior reduction maps onto the domain/task language of transfer learning (Pan & Yang, 2010; Ben-David et al., 2010), where a *domain* is a feature space with a marginal $P(X)$ and a *task* is a label space with a predictive $P(Y \mid X)$. Answer length is a property of the domain's $P(X)$ (which byte strings appear) and is specifiable independently of the task, i.e., of whether the model gets the answer right. Corpus BPB integrates the per-byte cost against each task's own empirical length distribution, so it adopts a different, self-selected $P(X)$ per task: this is the self-prior above. PRIORBPB instead fixes one explicit $P(X)$ (the shared length prior $s(k)$) for every task, so the reported quantity varies with the task's $P(Y \mid X)$ (continuation difficulty) and not with the domain's incidental length mix.

Two elementary structural facts. These reductions are consequences of two facts about any length-aggregated score, not of our particular estimator; both are weighted-average bookkeeping, with proofs in Appendix C. Write $\mathcal{A}_s[R] = \sum_k s(k)R(k)/\sum_k s(k)$ for the aggregation of a cost curve R under a position weight s, so that $\textsc{PriorBPB} = \mathcal{A}_s[R]$, and (Appendix C, Prop. 3) $\mathit{corpus}\textsc{BPB} = \mathcal{A}_{s^\star}[R]$ with $s^\star(k) = P_{\text{data}}(N \geq k)$ the eval set's own length-survival function, so corpus BPB and PRIORBPB are the same functional of the same curve, differing only in whether the position weight is read off the data or fixed in advance.

Proposition 1 (Weighted-average representation)**.** *Fix a byte budget n. Any score L on cost curves that is local (depends only on $R(1), \ldots, R(n)$), linear, monotone, and normalised ($L[\mathbf{1}] = 1$) equals $\mathcal{A}_s$ for a unique length prior $s \geq 0$ with $\sum_k s(k) = 1$. Consequently no non-degenerate score is length-invariant: if $\mathcal{A}_s[R]$ does not depend on the prior s, then R is constant on $1, \ldots, n$.*

Proposition 2 (Prior-robust comparison = pointwise dominance)**.** *For cost curves R_A, R_B on budget n, $\mathcal{A}_s[R_A] \leq \mathcal{A}_s[R_B]$ for every prior s if and only if $R_A(k) \leq R_B(k)$ for all $k \leq n$. Hence a model comparison flips between two priors only if the two cost curves cross.*

Proposition 1 says the only design freedom in a length-aggregated metric is the prior, and that corpus BPB (which reads the weight off each task's own length mix) is itself a choice, with no special claim to being the neutral default; the second half says one cannot escape a length prior without throwing away all per-position information. Proposition 2 says a comparison is prior-independent exactly when it reflects pointwise per-byte dominance, so any prior-sensitive ranking (including the scale 'anomaly' of Section 4) signals crossing cost curves rather than a robust capability gap. Proofs, the master identity, and the remaining reductions (monotonicity, flat-cost, the geometric closed form, and the large-budget limit) are in Appendix C.

Global vs. per-task prior. Fitting the prior per task, as Bayesian accuracy fits its slope b per model–dataset, only ranks answers within a task, whereas a cross-task per-byte scalar needs the opposite. By the reduction above, a task-local prior ($\mu =$ each task's own median) re-derives $\alpha/\bar{n} + \beta$ and puts the confound back. Across all 19 non-label tasks, Spearman(metric, mean length) is -0.78 for corpus BPB ($\mathrm{BF}_{10} \approx 8 \times 10^2$ at task level, ≈ 25 at the 12-family effective count: extreme thinning to strong, never absent) and -0.74 for a per-task prior ($\mathrm{BF}_{10} \approx 2.3 \times 10^2$, ≈ 13 family-adjusted; it collapses

back), whereas the shared prior gives -0.34 with a 95% credible interval $[-0.67, +0.04]$. We do not lean on a point-null Bayes factor here; for this near-null it is weak and prior-dependent (BF_{01} between 0.9 and 2.9 as σ_0 ranges over $[0.3, 2]$; Appendix H). Instead we test the main claim directly with a paired contrast: resampling tasks jointly, the shared-prior estimator is less length-confounded than corpus BPB ($\Delta|\rho| = 0.41$, 95% CrI $[0.15, 0.79]$) with posterior probability ≥ 0.999, and this survives a task-family clustered bootstrap. The residual -0.34 is itself a function of the prior mean (from -0.01 at $\mu = 40$ to -0.41 at $\mu = 160$) while the cross-task ordering stays stable (Spearman ≥ 0.84 for $\mu \geq 60$ vs. the frozen μ), so we report μ and read PRIORBPB as substantially, though not perfectly, length-neutral.

A diagnostic slice. The truncation-based slice $\textsc{PrefixBpB}(K) = \mathrm{mean}_i\, C_i(K)/K$ over items with $n_i \geq K$, together with the entry-cost/continuation-rate split $C(K) = [A(k_0) + R(k_0, K)(K - k_0)]$, localises where a model pays its bits. It is length-invariant by construction but estimable only where $n_i \geq K$ (Fig. 11); on one- or two-byte letter-label tasks no estimator can recover a byte-32 continuation rate.

C PROOFS OF THE STRUCTURAL RESULTS

Setup and notation. Fix a byte budget $n \geq 1$. A *cost curve* is a map $R : \{1, \ldots, n\} \to \mathbb{R}_{\geq 0}$; $R(k)$ is the marginal per-byte cost at content-byte position k (Eq. equation 2). A *position prior* (or weight) is $s : \{1, \ldots, n\} \to \mathbb{R}_{\geq 0}$ with $\sum_{k=1}^{n} s(k) > 0$. The aggregation functional is

$$\mathcal{A}_s[R] \;=\; \frac{\sum_{k=1}^{n} s(k)\, R(k)}{\sum_{k=1}^{n} s(k)}.$$

Write e_j for the point mass at position j ($e_j(k) = 1$ if $k = j$, else 0); it is a valid prior for $j \leq n$ and $\mathcal{A}_{e_j}[R] = R(j)$. For an eval set of items i with content lengths $n_i \leq n$ and per-byte costs $c_i(k)$, the length-survival function is $s^\star(k) = \#\{i : n_i \geq k\}$ and the empirical marginal curve is $R(k) = \big(\sum_{i:\, n_i \geq k} c_i(k)\big)/s^\star(k)$.

Elementarity. Both structural results below are elementary. Proposition 1 is the finite-dimensional representation of a positive, normalised linear functional (a weighted average) and Proposition 2 is the standard duality of the nonnegative cone; both proofs reduce to evaluating the functional at point masses (Rockafellar, 1970). We state them because the axioms, and the identification of $\mathcal{A}_s$ as the object of interest, are what the argument rests on rather than the derivations, and because an axiomatic characterisation is the standard way to show that a metric design choice is forced rather than derived ad-hoc or through heuristics/conventions (Krantz et al., 1971). Proposition 2 is pointwise dominance of two fixed cost curves; it is not Blackwell informativeness (Blackwell, 1953), which orders experiments.

Proposition 3 (Master identity)**.** $\mathit{corpus}\mathrm{BPB} \;=\; \dfrac{\sum_i \sum_{k=1}^{n_i} c_i(k)}{\sum_i n_i} \;=\; \mathcal{A}_{s^\star}[R]$.

Proof. Exchanging the order of summation (each item contributes to positions $k \leq n_i$), $\sum_i \sum_{k=1}^{n_i} c_i(k) = \sum_{k=1}^{n} \sum_{i:\, n_i \geq k} c_i(k) = \sum_{k=1}^{n} s^\star(k) R(k)$, the last step by the definition of $R(k)$. By the layer-cake identity, $\sum_i n_i = \sum_{k=1}^{n} \#\{i : n_i \geq k\} = \sum_{k=1}^{n} s^\star(k)$. Dividing gives $\mathcal{A}_{s^\star}[R]$. □

Proposition 3 is the sense in which corpus BPB and PRIORBPB are the same functional of the same curve: corpus BPB is $\mathcal{A}_s$ with the prior $s^\star$ read off the task's own lengths, whereas PRIORBPB fixes s in advance.

Proof of Proposition 1. Set $s(k) := L[e_k]$ for $k \leq n$. Linearity gives $L[\mathbf{0}] = 0$, and since $\mathbf{0} \leq e_k$ pointwise, monotonicity gives $s(k) = L[e_k] \geq L[\mathbf{0}] = 0$. For any curve R, the curves R and $\sum_{k=1}^{n} R(k)\, e_k$ agree at every position $\leq n$, so locality and linearity give

$$L[R] = L\Big[\textstyle\sum_{k=1}^{n} R(k) e_k\Big] = \sum_{k=1}^{n} R(k)\, L[e_k] = \sum_{k=1}^{n} s(k) R(k).$$

Taking $R = \mathbf{1}$ and using normalisation, $\sum_{k=1}^{n} s(k) = L[\mathbf{1}] = 1$; hence $L[R] = \sum_k s(k)R(k) = \mathcal{A}_s[R]$. The weight is unique among normalised priors: any representing s with $\sum_k s(k) = 1$ satisfies $s(k) = \mathcal{A}_s[e_k] = L[e_k]$. For the non-invariance clause, probe a fixed curve R with the priors e_j and e_k ($j, k \le n$): $\mathcal{A}_{e_j}[R] = R(j)$ and $\mathcal{A}_{e_k}[R] = R(k)$. If $\mathcal{A}_s[R]$ is independent of s then $R(j) = R(k)$ for all j, k, i.e., R is constant. □

Proof of Proposition 2. ($\Leftarrow$) If $R_A(k) \le R_B(k)$ for all $k \le n$ then, for any prior $s \ge 0$, $\sum_k s(k)R_A(k) \le \sum_k s(k)R_B(k)$; dividing by $\sum_k s(k) > 0$ gives $\mathcal{A}_s[R_A] \le \mathcal{A}_s[R_B]$. ($\Rightarrow$) Probing with $s = e_k$ gives $R_A(k) = \mathcal{A}_{e_k}[R_A] \le \mathcal{A}_{e_k}[R_B] = R_B(k)$ for each $k \le n$. □

Corollary 1 (Rank flips are crossings)**.** *If two priors s, t disagree ($\mathcal{A}_s[R_A] < \mathcal{A}_s[R_B]$ but $\mathcal{A}_t[R_B] < \mathcal{A}_t[R_A]$) then the curves cross: there exist $j, k \le n$ with $R_A(j) < R_B(j)$ and $R_B(k) < R_A(k)$.*

Proof. If $R_A \le R_B$ pointwise, Proposition 2 would give $\mathcal{A}_t[R_A] \le \mathcal{A}_t[R_B]$, contradicting the second inequality; so some k has $R_B(k) < R_A(k)$. Symmetrically, $R_B \le R_A$ pointwise would contradict the first inequality, so some j has $R_A(j) < R_B(j)$. □

Proposition 4 (The structural results do not rely on degenerate priors)**.** *Proposition 1's non-invariance clause and Proposition 2 both hold verbatim when 'every prior' is restricted to strictly positive priors.*

Proof. For $\varepsilon > 0$ the perturbed prior $e_j^{\varepsilon}(k) := e_j(k) + \varepsilon$ is strictly positive, with $\mathcal{A}_{e_j^{\varepsilon}}[R] = \big(R(j) + \varepsilon \sum_k R(k)\big)/(1 + \varepsilon n)$. Equating this for j and k (at $\varepsilon = 1$) forces $R(j) = R(k)$, giving non-invariance. For dominance, $\mathcal{A}_{e_k^{\varepsilon}}[R_A] \le \mathcal{A}_{e_k^{\varepsilon}}[R_B]$ reduces to $R_A(k) + \varepsilon \sum_j R_A(j) \le R_B(k) + \varepsilon \sum_j R_B(j)$; letting $\varepsilon \to 0^+$ yields $R_A(k) \le R_B(k)$. □

Proposition 5 (Tilt identity)**.** *For priors s, t with normalisations $\hat{s} = s/\sum s$, $\hat{t} = t/\sum t$, $\mathcal{A}_s[R] - \mathcal{A}_t[R] = \sum_{k=1}^{n} (\hat{s}(k) - \hat{t}(k))\, R(k)$: all disagreement between two length-aggregated scores is a reallocation of position mass.*

Proof. Immediate from $\mathcal{A}_s[R] = \sum_k \hat{s}(k)R(k)$ and $\mathcal{A}_t[R] = \sum_k \hat{t}(k)R(k)$. □

Reading the within-task residual. The identity also explains why the less biased within-task tilt (within-task length invariance, Section 6) sits slightly above the unbiased value (1.10 at 0.6B, decaying to 1.04 at 4B) rather than exactly at it: fixing a length prior reallocates a little too much mass towards long positions relative to the endogenous corpus weight, so the residual is a mild over-correction in the direction opposite to corpus BPB's under-costing of long golds (0.43), roughly $5\times$ smaller in magnitude and shrinking with scale, i.e., a small residual rather than a genuine reversal of the bias.

Proposition 6 (Monotonicity)**.** *Fix a prior s. If every per-byte cost weakly increases ($c_i(k) \to c_i'(k) \ge c_i(k)$), the empirical curve rises pointwise ($R' \ge R$) and hence $\mathcal{A}_s[R] \le \mathcal{A}_s[R']$: making any gold byte more surprising can never lower the reported score.*

Proof. $R(k)$ is a nonnegative-weighted mean of the $c_i(k)$, so $c_i' \ge c_i$ gives $R'(k) \ge R(k)$ for every k; apply the ($\Leftarrow$) direction of Proposition 2 with the fixed prior s. □

Proposition 7 (Flat-cost invariance)**.** *If $R \equiv a$ is constant then $\mathcal{A}_s[R] = a$ for every prior s.*

Proof. $\mathcal{A}_s[R] = \big(\sum_k s(k)\, a\big)/\sum_k s(k) = a$. □

Proposition 8 (Construct validity in the large-budget limit)**.** *If per-model curves converge pointwise, $R_m(k) \to h(k)$ for each $k \le n$, then for a fixed prior s, $\mathcal{A}_s[R_m] \to \mathcal{A}_s[h]$. If h is non-constant, $\mathcal{A}_s[h]$ depends on s (Proposition 1); so only a fixed prior identifies a length-free limit, whereas the endogenous corpus weight $s^\star$ gives different limits under different eval length laws.*

Proof. $\mathcal{A}_s[\cdot]$ is a finite weighted sum, hence continuous in the values $R(1), \ldots, R(n)$; convergence follows. The second claim is the non-invariance clause of Proposition 1. □

Proposition 9 (Geometric prior, affine cost)**.** *For an affine cumulative cost $C(k) = \alpha + \beta k$ (marginal $R(1) = \alpha + \beta$, $R(k) = \beta$ for $k \geq 2$) and the geometric prior $s(k) = (1 - 1/\mu)^{k-1}$ with $\mu > 1$,*

$$\mathcal{A}_s[R] = \beta + \frac{\alpha}{\mu\big(1 - (1 - 1/\mu)^n\big)} \xrightarrow[n\to\infty]{} \beta + \frac{\alpha}{\mu}.$$

Proof. Only the first byte incurs the fixed cost, so $\sum_{k=1}^{n} s(k)R(k) = \beta \sum_{k=1}^{n} s(k) + \alpha\, s(1)$, giving $\mathcal{A}_s[R] = \beta + \alpha\, s(1)/\sum_{k=1}^{n} s(k)$. Here $s(1) = 1$ and, by the finite geometric sum, $\sum_{k=1}^{n}(1 - 1/\mu)^{k-1} = \mu\big(1 - (1 - 1/\mu)^n\big)$. Since $0 \leq 1 - 1/\mu < 1$, the truncation factor $1 - (1 - 1/\mu)^n \to 1$, giving the limit. □

Proposition 10 (Corpus identity, Eq. equation 1)**.** *Let (α, β) be the ordinary-least-squares fit of total bits L_i on content length n_i over m items. Then $\big(\sum_i L_i\big)/\big(\sum_i n_i\big) = \alpha/\bar{n} + \beta$.*

Proof. The first normal equation is $\sum_i (L_i - \alpha - \beta n_i) = 0$, i.e., $\sum_i L_i = m\alpha + \beta \sum_i n_i$. Dividing by $\sum_i n_i = m\bar{n}$ gives $\alpha/\bar{n} + \beta$. (Only the first normal equation is used; no assumption on the length variance is needed.) □

Proposition 11 (Byte attribution)**.** *Linear within-token attribution (each token's bits shared equally across the bytes it spans, a zero-byte token folding onto the previous byte) is conservative: the attributed sequence has one entry per gold byte (its length equals the total token byte-length) and, for any item with at least one byte, its entries sum to the total token bits. Removing the leading-space offset leaves the n_i content bytes, and corpus* BPB *on them is total content bits over total content bytes (Prop. 3).* PRIORBPB *keeps all token bits in the numerator (including the dropped leading-space token) while using content bytes in the denominator, so it exceeds the content corpus* BPB *by exactly $\big(\sum_i$ leading-space bits$_i\big)/\big(\sum_i n_i\big)$, which is zero iff every offset is* 0*.*

Proof. Attribution reassigns bits within each token and emits one entry per spanned byte, so it preserves both the total byte count and the total bit mass. Splitting the attributed sequence at the offset, its bits decompose as leading-space bits plus content bits; the content ratio is corpus BPB by Proposition 3. Writing the PRIORBPB numerator (all token bits) as content bits plus leading-space bits over the common content-byte denominator gives the stated additive correction, which vanishes exactly when no leading-space byte is removed. □

D THE SAME PANEL SET ON QWEN3-0.6B AND OLMO-3-7B

Figure 2 in the main text uses Qwen3-4B so the intercept on GSM8K is large ($\alpha = 153$ bits) and the confound is not a small-model artefact. The affine split, the SQuAD hyperbola, and the heterogeneous length shares are the same shape at 0.6B and on OLMo-3-7B-Instruct (Figs. 4 and 5). What changes is the mix: at 0.6B the GSM8K intercept is only a seventh of mean bits, while on OLMo-3-7B it is almost absent ($\alpha = 4$ bits) and length accounts for nearly all of that task's cost.

E THE CASE STUDY IN FULL

This expands the condensed account of Section 4. The (α, β) attribution is fit-contingent (§2), but the direction is clear and corroborated by the raw per-token traces. The fixed cost α grows with scale on 19 of 19 tasks (e.g., GSM8K $22 \to 93 \to 153$ bits; NaturalQuestions $12 \to 22 \to 34$), while the per-byte rate β shows no such trend (Fig. 6). Per-token traces put the cost where the format-mismatch mechanism predicts: the GSM8K position-1 bits run $7.4 \to 19.1 \to 22.8$ across the ladder while the later-token plateau barely moves (Fig. 9).

A controlled replicate. We re-ran all three models on all tasks with `enable_thinking=False`. The flag changes the prompt on only 8 of 19 tasks (the rest are byte-identical and thus uninformative); on those 8 the anomaly strengthens: 0.6B wins on 18 of 19 tasks with thinking off. So the apparent inverse scaling is a fixed-cost/format effect BPB cannot separate, not a statement about model capability.

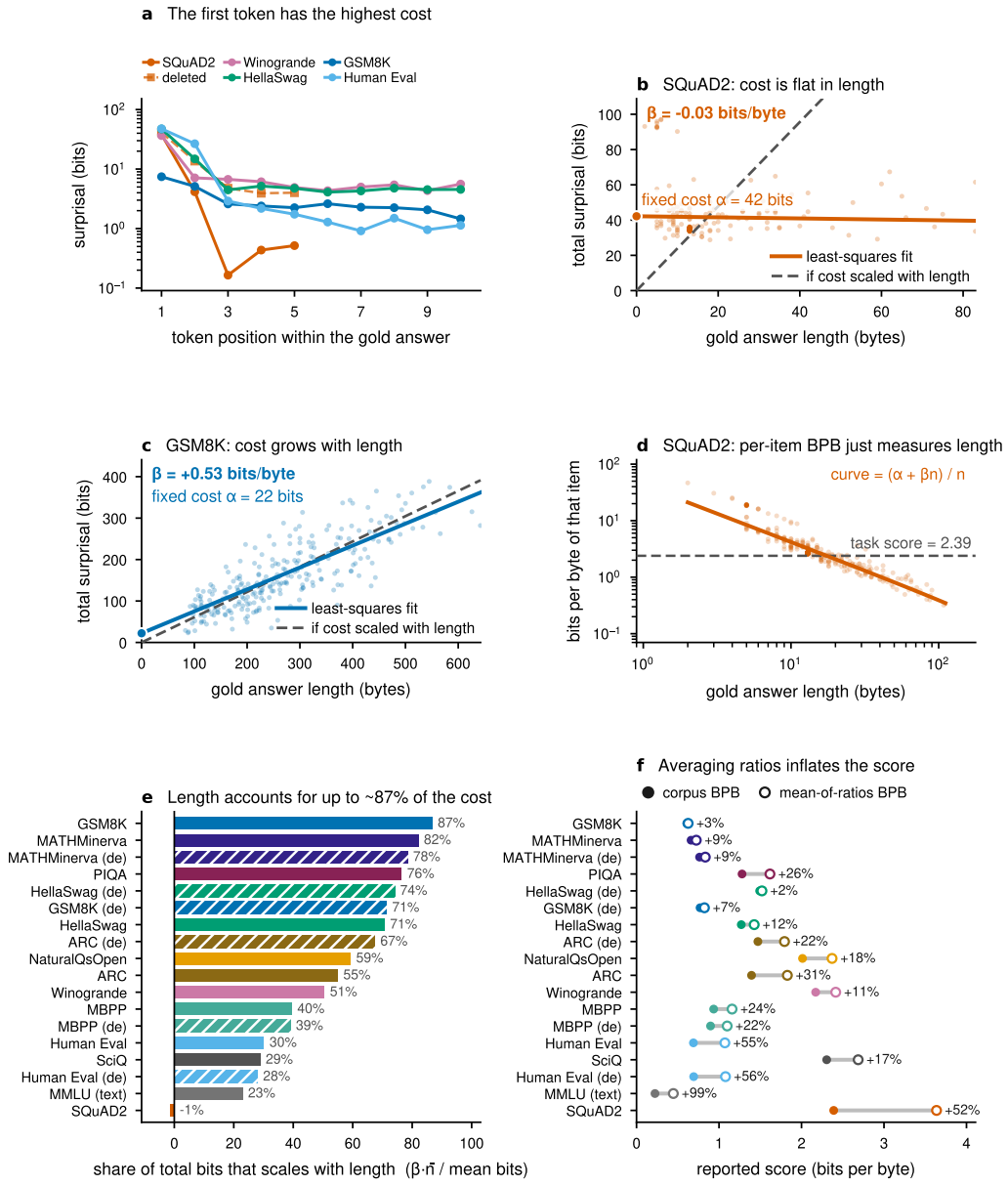


Figure 4: Same six panels as Fig. 2, for Qwen3-0.6B. SQuAD is a nearly pure intercept ($\alpha = 42$ bits, $\beta \approx 0$); GSM8K mixes a small intercept with a slope ($\alpha = 22$, $\beta = +0.53$). Length accounts for up to approximately 87% of the cost; mean-of-ratios inflates the score by up to 99%.

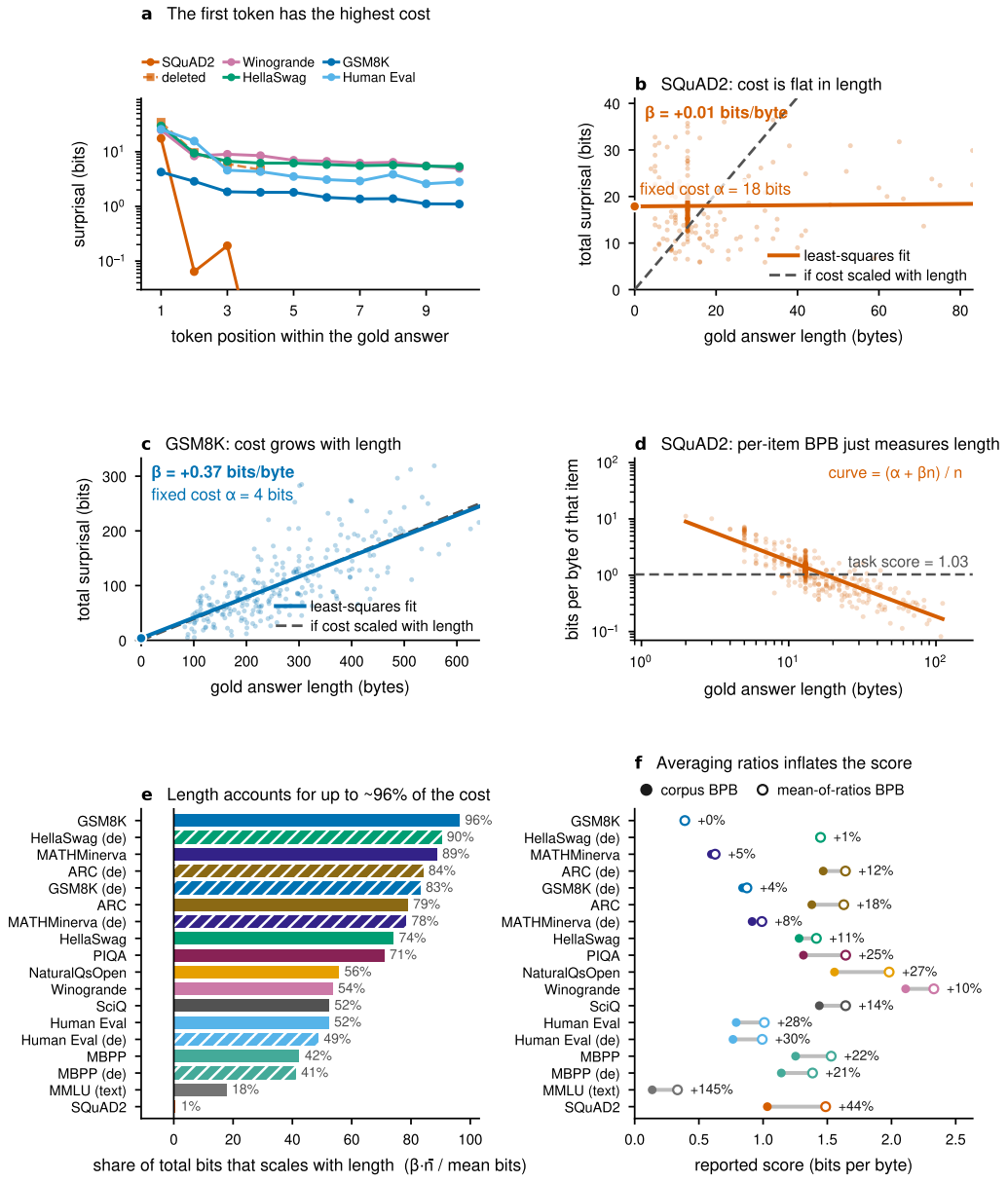


Figure 5: Same six panels as Fig. 2, for OLMo-3-7B-Instruct. SQuAD is again a pure intercept ($\alpha = 18$ bits, $\beta \approx 0$); GSM8K is almost pure slope ($\alpha = 4$, $\beta = +0.37$). Length accounts for up to approximately 96% of the cost; mean-of-ratios inflates the score by up to 145%.

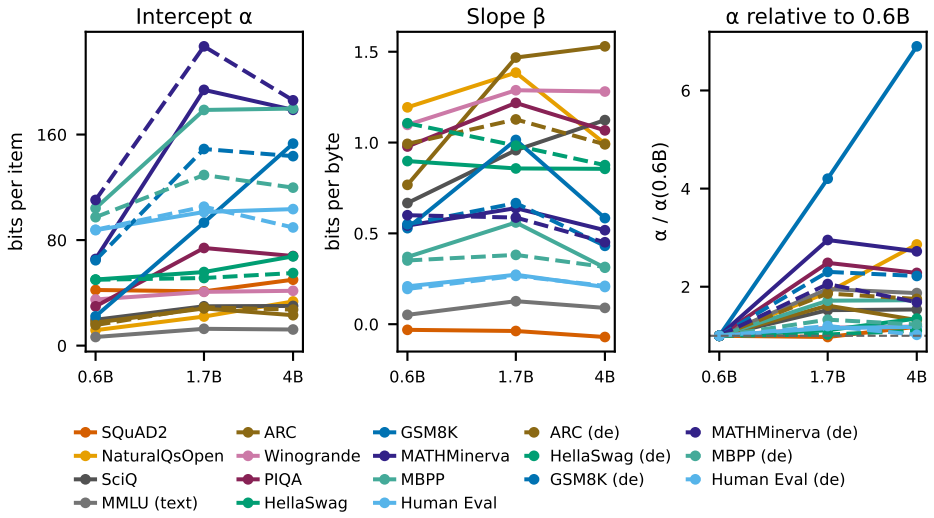


Figure 6: The scale anomaly is an intercept effect. The fixed cost α grows with model size on 19/19 tasks (left), whereas the per-byte rate β does not (right). Corpus BPB mixes the two via Eq. equation 1, so a growing α shows up as 'bigger is worse'.

The compounding aggregation error, per task. The mean-of-ratios aggregator obeys $\mathbb{E}[L/n] \approx \alpha\,\mathbb{E}[1/n] + \beta$, and since $\mathbb{E}[1/n] \geq 1/\bar{n}$ it includes a strictly larger fixed-cost term than corpus BPB. Here it inflates the score by 2–99% (Full-Text MMLU $+99\%$, SQuAD $+52\%$, HumanEval $+55\%$; GSM8K $+3\%$) and is $\sim$13% noisier at every sample size, which is why the checklist's first item is to report corpus BPB rather than mean-of-ratios BPB.

F Additional figures

G Further properties

The remaining checks hold by construction or by wide margins. The estimator is *monotone*, making any token more surprising never lowers the score (0/19 violations; guaranteed by non-negative weights, Prop. 6). It is *statistically efficient*, with estimator noise comparable to corpus BPB (CV ratio 0.97). It is *cross-lingually neutral*: because token counts for the same content differ markedly across languages (Petrov et al., 2023; Ahia et al., 2023), on matched German/English pairs the median $|\text{DE/EN} - 1|$ improves from 0.100 to 0.027, and on Qwen3 bits-per-byte is closer to language-neutral than bits-per-token on 8/12 gold-translated measurements (0.14 vs. 0.21; Fig. 10). This byte-over-token advantage is tokeniser-specific and does not hold universally: on a second tokeniser (SmolLM2 (Ben Allal et al., 2025)) the byte denominator is actually the less language-stable of the two (Appendix I), so we claim byte-normalisation helps here rather than everywhere. Finally, the attribution is *robust*: linear vs. token-boundary byte attribution agree to 10% and the leading-space convention moves the score by 4% (Prop. 11).

H Robustness of the Bayesian analysis

Inference protocol (full). We report Bayesian summaries throughout, rather than p-values, because a credible interval and a posterior probability directly express how probable an effect is given the data. (Our estimator likewise aggregates against an explicit length prior; but it is a prior-weighted average, and we do not claim it is a posterior or otherwise an object of Bayesian inference.) Rank correlations and the within-task tilt are summarised by their full posterior via the Bayesian bootstrap (Rubin, 1981; Gelman et al., 2013) (10^5 Dirichlet$(1, \ldots, 1)$ resamples that reweight a fixed-rank weighted Pearson correlation), giving 95% credible intervals and probabilities such as $P(\rho < 0)$. For the point null $\rho = 0$ we compute a Fisher-z Savage–Dickey Bayes factor (Ly et al., 2016) on the Jeffreys scale

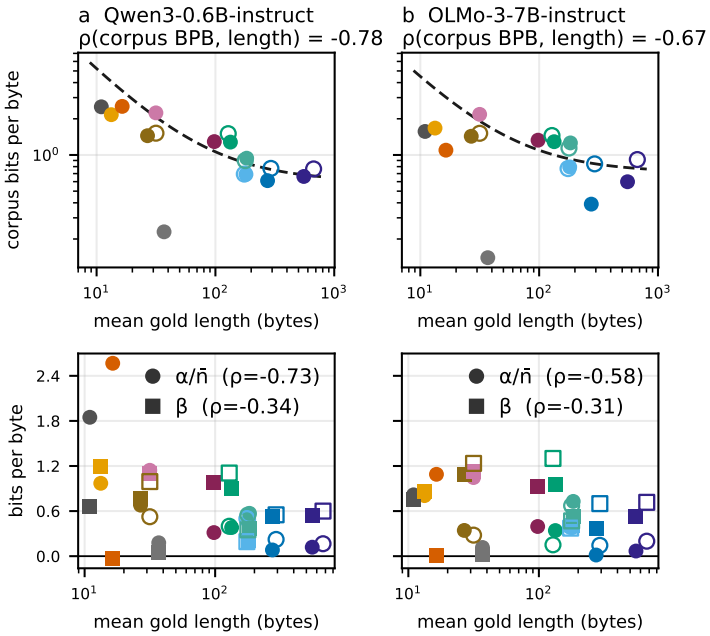


Figure 7: BPB vs. mean gold length across the same 19 length-varying tasks (log-x). Left: Qwen3-0.6B-instruct. Right: OLMo-3-7B-instruct. The reported score falls as golds lengthen on both families (Spearman -0.78 and -0.67 for corpus BPB; -0.83 and -0.79 for mean-of-ratios BPB), and the dependence sits in the fixed-cost term $\alpha/\bar{n}$ (-0.73 / -0.58) rather than the per-byte rate β (-0.34 / -0.31). The cross-task ordering is largely a length ordering. Filled markers are English tasks and open markers their matched German counterparts; in the lower panels circles are the fixed-cost term $\alpha/\bar{n}$ and squares the per-byte rate β. Per-task colours match those in Fig. 2(e).

(Jeffreys, 1961) (> 10 strong, > 100 extreme); because the tasks are not independent (EN/DE pairs, MC variants) we report it at both the task count $n_{\text{eff}} = 19$ and the task-family count 12, and (being a Pearson-on-ranks approximation) use it only as point-null evidence, defaulting primary claims to the assumption-light bootstrap posterior. We report posteriors directly rather than thresholding against a region of practical equivalence; where a scale for 'negligible' is helpful we cite the conventional $|\rho| < 0.1$ band (Kruschke, 2018) only as a reference, not a decision rule. Every headline is reported under both the plain and a task-family clustered bootstrap, with the prior width $\sigma_0 \in [0.3, 2]$ swept and a paired contrast for the main 'the shared prior does the work' claim (Appendix H), so no conclusion rests on the exchangeability of the 19 tasks or a single prior.

We check the two main inferences against three objections: a mis-specified prior, non-independent tasks, and an arbitrary shared-prior mean μ.

Prior sensitivity. Sweeping the Fisher-z prior width $\sigma_0 \in \{0.3, 0.5, 1, 1.5, 2\}$ (Fig. 12a), the corpus length dependence stays extreme throughout (BF_{10} from 104 to 812, always ≥ 100); mean-of-ratios BPB (the number evaluation frameworks actually report) is if anything *more* confounded ($\rho = -0.83$ vs. corpus -0.78; BF_{10} from 469 to $\sim$9000, and still 'very strong' at $\text{BF}_{10} \approx 91$ under the n_{eff}=12 family count), so the debiasing corrects a genuine defect in the default metric rather than a strawman. The shared-prior estimator never provides more than moderate evidence for the null (BF_{01} from 0.9 at σ_0=0.3 to 2.9 at σ_0=2, mostly inconclusive), which is why we do not present it as an established equivalence and instead rely on the paired contrast below. The paired contrast against mean-of-ratios BPB is itself strong: $\Delta = |\rho_{\text{mor}}| - |\rho_{\text{shared}}| = 0.46$ (95% CrI $[0.21, 0.83]$; 0.47, $[0.18, 0.84]$ clustered), with posterior probability > 0.999 that the shared prior is the less length-confounded of the two.

Effective sample size. The Bayes factors above use the Fisher-z likelihood variance $1/(n_{\text{eff}} - 3)$ with $n_{\text{eff}} = 19$ tasks. Because the 19 tasks are not independent (§2), we also recompute them at

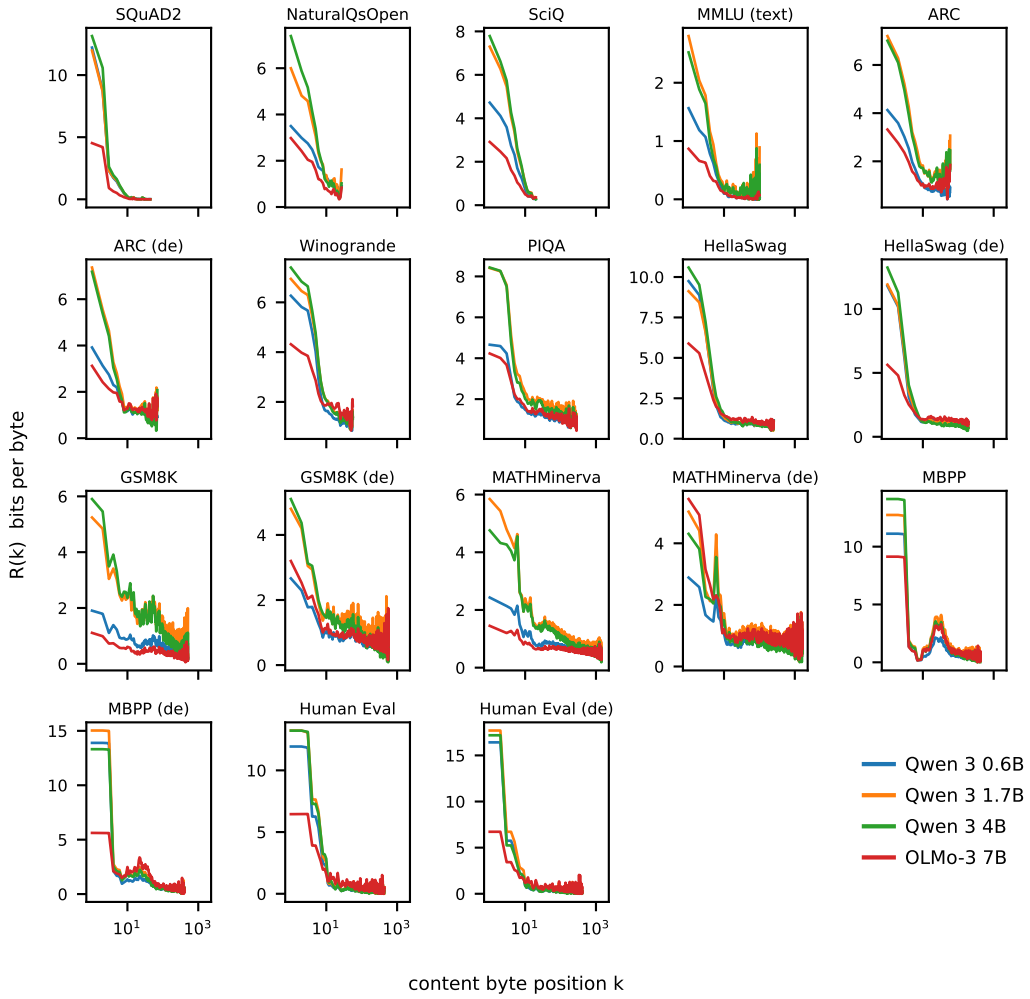


Figure 8: The primitive $R(k)$: marginal per-byte cost by content-byte position, estimated per (model, task) using every byte of every gold, for the three Qwen3 scales (0.6B/1.7B/4B) and OLMo-3 7B. Task-specific length–cost relationships are captured here; the shared prior of Eq. equation 3 only sets how positions are weighted into a comparable scalar.

the task-family count $n_{\text{eff}} = 12$: the corpus length dependence reduces from extreme to strong but never vanishes ($\text{BF}_{10} \approx 22$ for the Section 3 length correlation and ≈ 25 for the cross-task deconfounding, both ≥ 10), the per-task-prior control stays strong (≈ 13), and the shared-prior estimator remains inconclusive ($\text{BF}_{10} \approx 0.53$). The qualitative split (corpus confounded, shared prior not distinguishable from neutral) is unchanged, so no conclusion rests on treating the 19 tasks as independent draws.

Non-exchangeable tasks. The 19 tasks collapse to 12 families once EN/DE pairs and MC variants are grouped. Under a task-family clustered Bayesian bootstrap the corpus credible interval is essentially unchanged ($[-0.92, -0.46]$) while the shared-prior interval widens to $[-0.67, +0.12]$ with $P(\rho < 0) = 0.92$, i.e., acknowledging the dependence only strengthens the qualitative conclusion (corpus confounded, shared prior not distinguishable from neutral).

Paired contrast. Resampling tasks jointly and comparing $\Delta = |\rho_{\text{corpus}}| - |\rho_{\text{shared}}|$ gives $\Delta = 0.41$ (95% CrI $[0.15, 0.79]$, plain; $\Delta = 0.43$, $[0.14, 0.81]$, clustered) with posterior probability ≥ 0.999 that the shared prior is the less length-confounded of the two. This paired quantity, not two separate Bayes factors, is the evidence for 'the shared prior does the work.' Because the shared-prior posterior

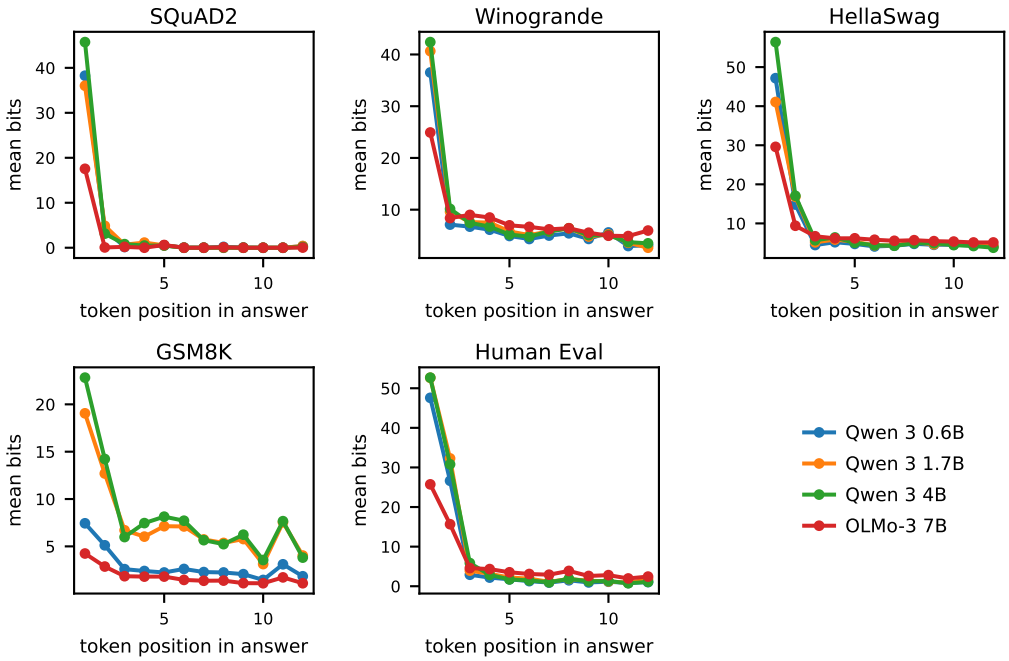


Figure 9: Mean surprisal by gold-token position across 5 tasks, for the three Qwen3 scales (0.6B/1.7B/4B) and OLMo-3 7B. The first token dominates; SQuAD's later tokens are near-free when the answer is in context. For Qwen3 the GSM8K position-1 spike grows with model scale ($7.4 \to 19.1 \to 22.8$ bits), matching the intercept growth of Section 4.

straddles zero, $|\rho_{\text{shared}}|$ is folded at the origin, which makes Δ conservative (it cannot borrow strength from a sign it does not have); the > 0.999 mass is thus a lower bound on the separation.

Shared-prior mean. Varying μ (Fig. 12b) trades residual length correlation (-0.01 at μ=40 to -0.42 at μ=180) against how much late-answer content is weighted, but the induced cross-task ordering is stable across sensible prior means (Spearman ≥ 0.84 for $\mu \geq 60$ against the frozen μ=111; only the extreme μ=40 falls to 0.60). That value sits well below the pooled median answer length and below the typical gold of almost every task, so it down-weights content the tasks genuinely produce and is outside the sensible range; for $\mu \geq 60$ the ordering is stable. We therefore report μ explicitly and treat PRIORBPB as substantially length-neutral, though not perfectly so.

Is μ=111 arbitrary? Marginalising it out. μ is not a free knob: it is the pooled cross-task median content length. To show no conclusion rides on its point value we place a hyperprior on it and integrate. Under a lognormal centred at 111 (log-scale σ=0.3, central 95% $\approx [60, 200]$) the μ-marginal posterior for the shared-prior length correlation is -0.36 (95% CrI $[-0.69, +0.08]$, $P(\rho < 0) = 0.95$); a uniform$[60, 180]$ hyperprior gives -0.37 ($[-0.69, +0.09]$). Both are far milder than corpus BPB and straddle zero, so the qualitative split survives marginalising away the choice of μ.

Survivorship. Because $R(k)$ at large k is estimated only on the long-answer survivors, one might worry the neutrality is an artefact of which items survive. Two isolations say otherwise. Imposing a content-byte floor (dropping short items outright) leaves the shared-prior length correlation at -0.34 for floors up to 8 bytes (corpus stays ≤ -0.76), so short items are not what neutralises it. Conversely, capping the integrated budget at $k_{\max}$ bytes (which removes the sparsely-supported deep-tail positions) drives the shared-prior correlation to ≈ 0 ($+0.09, +0.01, -0.04$ at $k_{\max} = 32, 64, 16$; paired advantage $+0.62$), locating the residual in the tail rather than in survivorship. PRIORBPB integrates the full curve (no cap) and thus reports the more conservative -0.34.

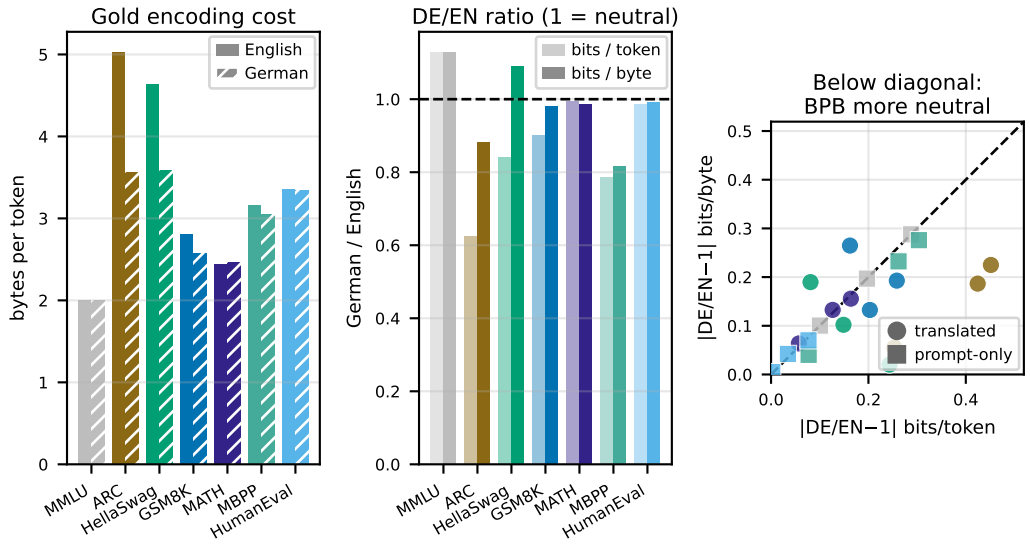


Figure 10: Cross-lingual neutrality check on matched German/English pairs. **Left:** gold encoding cost in bytes per token (higher = more efficient tokenisation); the English-centric tokeniser chops German into more, smaller pieces. **Middle:** German/English ratio for bits-per-token vs. bits-per-byte (1.0 = language-neutral). **Right:** deviation from neutrality; points below the diagonal mean bits-per-byte is more language-neutral than bits-per-token, shown separately for translated golds and prompt-only pairs (gold is Python). The byte denominator removes most of the bits-per-token gap. Broken out by estimator, the median $|\text{DE/EN} - 1|$ is 0.13 (corpus), 0.13 (mean-of-ratios), 0.11 (PREFIXBPB) and 0.18 (PRIORBPB): byte normalisation is close to language-neutral for all four, but the fixed length prior cedes a little cross-lingual stability, which is why this per-estimator comparison appears here rather than in the main length figure (Fig. 3).

Token-level attribution. PRIORBPB splits a straddling token's bits byte-linearly. A production evaluation framework that cannot split tokens would instead attribute each token's bits at its boundary. This coarse variant changes the per-task score by a median 8.5% (95th pct 35%) and, if anything, is more length-coupled ($\rho = -0.49$ vs -0.34 for byte-linear), while preserving the cross-task ordering (Spearman 0.93), so byte-linear attribution is the better default but the cheap approximation is a usable fallback.

I EXTERNAL CONTROLS: BASE CHECKPOINTS AND A SECOND TOKENISER

Two objections to the case study of Section 4 are that the fixed-cost growth might be an artefact of scoring completion-style golds through an instruct chat template, and that the length identity might be specific to the Qwen3 tokeniser. We run two controls that hold the estimator fixed and vary exactly these factors.

Base checkpoints, no chat template. We re-score Qwen3-{0.6,1.7,4}B-Base (the same family and scales as the main matrix, but not instruction-tuned) with a plain concatenation formatter (no chat roles, no reasoning cue). This removes the instruct template entirely, leaving the standard base-model setup in which prompt and gold are simply concatenated. If the fixed cost α still climbs with scale here, the growth is a property of the measurement rather than the template (Fig. 13).

A second tokeniser. We re-score SmolLM2-{135M,360M,1.7B} (a different family with a non-Qwen tokeniser, also base checkpoints) on the same English tasks plus the matched German pairs. This re-verifies the corpus identity $corpus\text{BPB} = \alpha/\bar{n} + \beta$ on an independent tokeniser and repeats the cross-lingual bits-per-byte vs. bits-per-token comparison on it.

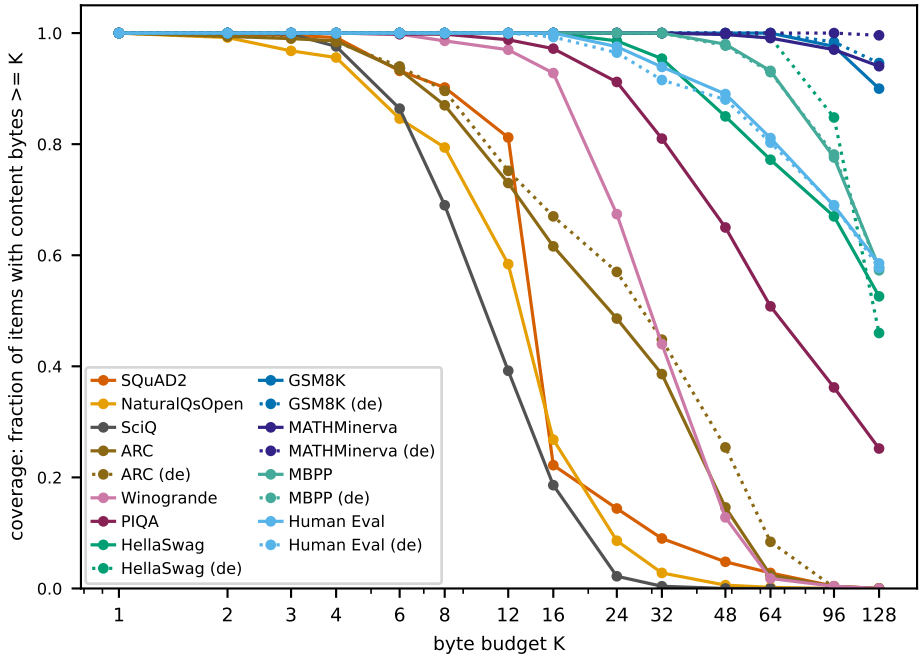


Figure 11: Coverage of the truncation-based diagnostic PREFIXBPB(K): fraction of items with content bytes $\geq K$. Short-gold tasks lose support quickly, which is why the headline scalar is the all-bytes PRIORBPB and K is used only as a diagnostic.

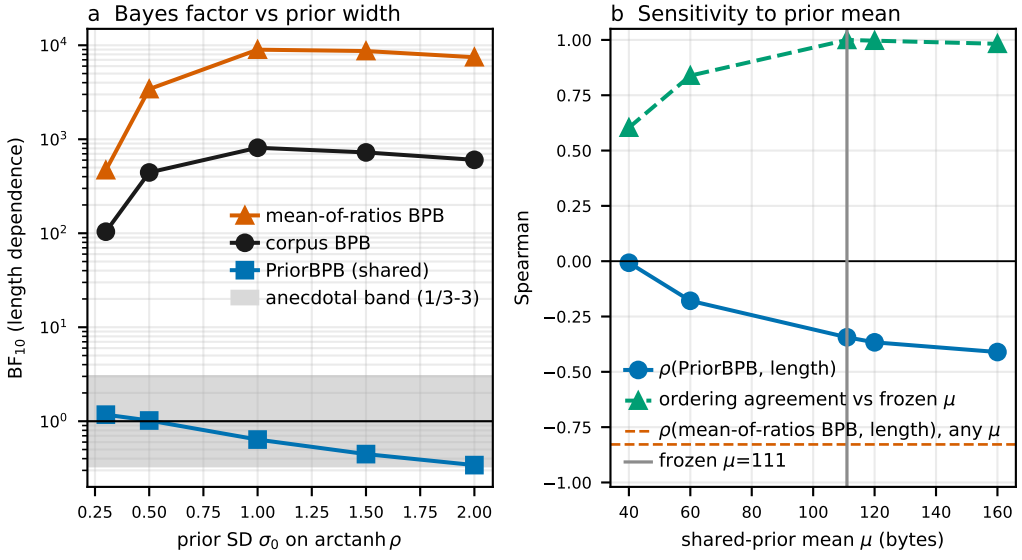


Figure 12: Robustness of the Bayesian analysis. **(a)** Bayes factor for length dependence vs. the Fisher-z prior width σ_0: mean-of-ratios BPB (the evaluation-framework default) and corpus BPB are both extreme for every σ_0 (mean-of-ratios BPB the more so) while the shared-prior estimator stays in/near the anecdotal band. **(b)** Sensitivity to the shared-prior mean μ: the residual length correlation grows with μ but stays well above the μ-independent mean-of-ratios baseline (dashed), and the cross-task ordering agrees with the frozen μ=111 at Spearman ≥ 0.84 for $\mu \geq 60$ (only the extreme μ=40 dips to 0.60).

Results. On the base checkpoints (concatenation formatter, no chat template) the fixed cost α climbs the full $0.6 \to 4$B ladder on 25% of 12 tasks, versus 100% of 19 for the instruct checkpoints, so the fixed-cost growth is attenuated once the instruct template is removed. For GSM8K, the base fixed cost runs $\alpha = 12/6/8$ bits across $0.6/1.7/4$B versus $\alpha = 23/94/154$ for instruct.

Identity and second tokeniser. The corpus identity $corpus\text{BPB} = \alpha/\bar{n} + \beta$ holds to $1.1e-15$ bits across every base and SmolLM2 run, confirming the decomposition is exact on an independent tokeniser. The cross-lingual comparison repeats on both tokenisers—SmolLM2 (median $|\log_2$ DE/EN$|$: 0.757 vs. BPT 0.229); Qwen3 (0.258 vs. BPT 0.417)—so the byte denominator lowers the DE/EN gap for Qwen3 but not SmolLM2: which scale is more language-stable is itself tokeniser-dependent, and the byte-normalisation advantage should be read as tokeniser-specific rather than universal.

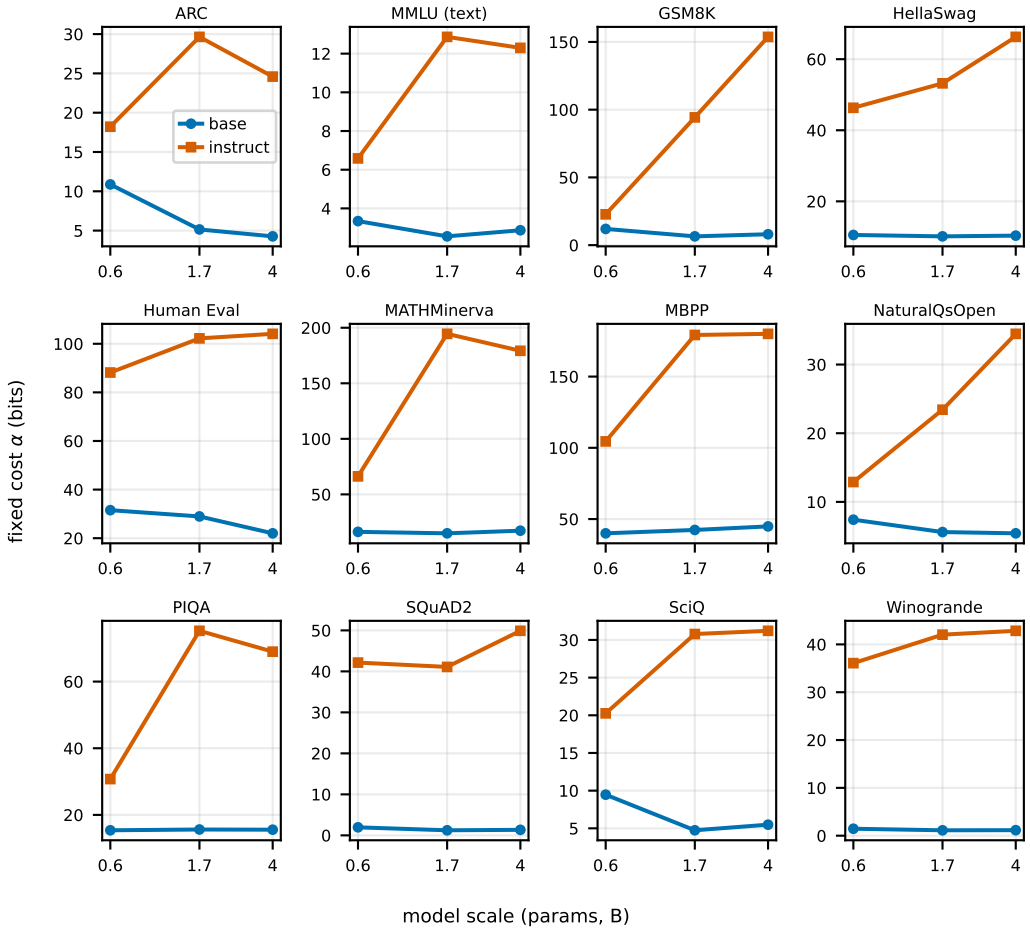


Figure 13: Fixed cost α (bits) vs. model scale for base checkpoints scored without a chat template (solid) and the instruct checkpoints (dashed), per task. The affine fit $L = \alpha + \beta n$ is taken over content bytes; $corpus\text{BPB} = \alpha/\bar{n} + \beta$ holds exactly. See §I text for the direction of the effect.

J Generalisation: second families, held-out transfer, within-suite CV

Generalisation across families and to held-out tasks. These length results are not specific to Qwen3-instruct. Re-scoring two further families with their own complete per-token matrices, the shared prior takes the cross-task length correlation from $-0.34/-0.48$ (corpus) to $+0.06$–$+0.09$ on SmolLM2-base (16 tasks, correlation essentially removed) and from $-0.45/-0.50$ to $-0.20/-0.32$

on Qwen3-Base (12 tasks); it is the less length-confounded estimator on all six additional checkpoints ($P \geq 0.76$, and ≥ 0.94 on Qwen3-Base; Appendix J). Freezing $\mu = 111$ (fit on the 22-task suite) and scoring a held-out suite of four datasets spanning 7–153 gold bytes (CommonsenseQA (Talmor et al., 2019), DROP (Dua et al., 2019), MedQA (Jin et al., 2021), GoldenSwag (Chizhov et al., 2025)), the mean length correlation drops in magnitude from -0.73 (corpus) to $+0.37$ (PRIORBPB) and the shared prior is the less length-confounded estimator on every model of both families ($P(\Delta > 0) = 1.000$), the debiasing transfers to held-out tasks the prior was not refit on. A within-suite cross-validation agrees: fitting μ on a random half of the tasks and testing on the other half, the held-out length correlation is -0.76 for corpus vs -0.32 for the shared prior (paired advantage $+0.45$, $P = 0.99$; Appendix J).

The main-text length numbers are Qwen3-instruct. Here we show the length-debiasing is neither single-family nor tuned to the 22-task suite.

Further families. We recompute the within-task tilt and the cross-task length correlation on the complete per-token matrices of SmolLM2-base (a different tokeniser, 16 tasks), Qwen3-Base (12 tasks), and OLMo-3-7B-Instruct (the same 19 length-varying tasks as the Qwen3-instruct analysis), with the same frozen $\mu = 111$ and the same family-clustered Bayesian bootstrap (Table 2). On SmolLM2 the shared prior largely removes the correlation ($-0.34/-0.48 \rightarrow +0.06–+0.09$); on Qwen3-Base it roughly halves it ($-0.45/-0.50 \rightarrow -0.20/-0.32$); on OLMo-3-instruct it all but removes it ($-0.67 \rightarrow -0.01$). The paired contrast places the shared prior below corpus in length dependence on all seven checkpoints ($P \geq 0.76$ on SmolLM2, ≥ 0.94 on Qwen3-Base, $P = 0.988$ on OLMo-3-7B-Instruct), and the within-task tilt moves from below-unbiased (0.46–0.75; 0.49 on OLMo) to essentially unbiased (1.10–1.22).

Table 2: Further-family length behaviour (frozen $\mu = 111$): the debiasing generalises beyond Qwen3-instruct. ρ is Spearman(score, mean content bytes); $\Delta = |\rho_{\text{corpus}}| - |\rho_{\text{shared}}|$ with $P(\Delta > 0)$ from a task-family-clustered bootstrap; tilt is the top/bottom length-quintile ratio (1 = unbiased). OLMo-3-instruct uses the same 19 length-varying tasks as the main analysis.

Family	Model	ρ_{corpus}	ρ_{shared}	Δ	$P(\Delta>0)$	tilt (corpus→prefix)
SmolLM2-base	135M	-0.34	$+0.07$	$+0.22$	0.77	$0.75 \rightarrow 1.19$
SmolLM2-base	360M	-0.34	$+0.09$	$+0.22$	0.76	$0.74 \rightarrow 1.17$
SmolLM2-base	1.7B	-0.48	$+0.06$	$+0.37$	0.93	$0.75 \rightarrow 1.22$
Qwen3-Base	0.6B	-0.50	-0.32	$+0.15$	0.95	$0.48 \rightarrow 1.10$
Qwen3-Base	1.7B	-0.46	-0.27	$+0.17$	0.95	$0.47 \rightarrow 1.13$
Qwen3-Base	4B	-0.45	-0.20	$+0.23$	0.94	$0.46 \rightarrow 1.15$
OLMo-3-instruct	7B	-0.67	-0.01	$+0.52$	0.99	$0.49 \rightarrow 1.20$

Out-of-suite transfer. We freeze $\mu = 111$ (fit on the 22-task suite) and score a held-out suite of four datasets: CommonsenseQA, DROP, MedQA (full-text cloze) and GoldenSwag, scored per-token on all three Qwen3-instruct and all three SmolLM2 checkpoints. Three contribute genuinely new content; GoldenSwag is a curated subset of HellaSwag (which is in the fitting suite), so it is held out only in that μ was not refit on it. Their mean gold lengths span 7 to 153 content bytes. Across the length-varying held-out tasks the mean length correlation is -0.73 for corpus BPB vs $+0.37$ for PRIORBPB; the shared prior is the less length-confounded estimator on every one of the six models and in both families ($\Delta = +0.37$ pooled, $P(\Delta > 0) = 1.000$). The residual flips sign and roughly halves in magnitude rather than reaching exactly zero on this small four-task set, consistent with the μ-sensitivity above; the main claim (unanimous paired advantage over corpus on held-out tasks the prior was not refit on) holds.

Within-suite cross-validation. As a higher-powered complement we split the 22-task suite into random halves (300 splits), fit μ as the pooled median content length on the train half, and evaluate on the held-out half. The fitted μ is stable (median 129) and its length-neutrality transfers: the test-half length correlation is -0.76 for corpus vs -0.32 for the shared prior, a paired advantage of $+0.45$ with $P(\text{shared less confounded}) = 0.993$.

K DOES PRIORBPB TRACK ACCURACY BETTER THAN MEAN-OF-RATIOS BPB?

The main result of this comparison (PRIORBPB tracks downstream accuracy reliably better than mean-of-ratios BPB on every stratum) is reported in the main text (Section 6, Table 1); this appendix gives the full design, set of models, and per-gold robustness behind it.

BPB is used as a smooth proxy because accuracy is noisy and quantised during pretraining. Whether a log-likelihood metric actually tracks downstream capability is contested: loss-to-accuracy scaling fits are often unreliable (Lourie et al., 2025), and the perplexity–downstream link can require task-specific adaptation to reappear (Zhang et al., 2026). This raises a proxy-quality question that length de-confounding does not answer: does the less biased estimator track downstream accuracy more faithfully than the standard one? We answer with a paired, hierarchical Bayesian comparison of PRIORBPB (geometric shared prior, frozen $\mu = 111$) against the mean-of-ratios aggregator that evaluation frameworks actually report (Gao et al., 2024).

Design. A *condition* is a (ladder, task, accuracy-gold) triple; a *ladder* is one monotone capability sequence. For each condition we compute a *fidelity*, how well a metric's movement mirrors accuracy's along the ladder, oriented so higher is better and BPB's lower-is-better sign is absorbed: primary $= -$Spearman, with $-$Pearson and pairwise concordance as robustness. The paired estimand is $\Delta_c = \text{fidelity}(\text{PRIORBPB}; c) - \text{fidelity}(\text{mean-of-ratios}; c)$, aggregated by a Bayesian bootstrap clustered on (model-family $\times$ task-family) into a posterior mean Δ, a 95% credible interval, and the direction probability $P(\Delta > 0)$. We read each result by its *direction* (the sign and $P(\Delta > 0)$) and its *magnitude* (the posterior mean and credible interval), rather than by thresholding Δ against an arbitrary practical-equivalence margin: a comparison is reliably signed when $P(\Delta > 0)$ is near 1 (or 0) and the credible interval excludes zero, and undetermined when the interval spans zero. As a control, PRIORBPB must not lose against a length-independent (letter-label) or a length-corrected (Bayesian (Oostermeijer, 2026)) accuracy gold; otherwise an apparent win could be a shared-length artefact.

Models. *During pretraining* (dense, clean ladders): Pythia (Biderman et al., 2023) $\{160\text{M}, 410\text{M}, 1\text{B}, 1.4\text{B}\}$ at up to 24 log-spaced steps (step$1000 \ldots 143000$; one tokeniser, shared data order), scored with a plain concatenation formatter, plus the Pythia-*deduped* ladders of the same sizes as an independent second data order. *End of pretraining* (cross-tokeniser breadth): the Qwen3-Base, SmolLM2, and Qwen3-instruct runs of the main paper, plus Qwen2.5-$\{0.5, 1.5\}$B (Yang et al., 2024) and OLMo-2-1B (Team OLMo et al., 2024) and their instruct variants (Llama-3.2 is gated and dropped out). Tasks: cloze/full-text ARC, HellaSwag, PIQA, SciQ, WinoGrande, MMLU and letter-label ARC/MMLU/PIQA/SciQ (all emitting accuracy and per-token bits in one pass).

Informative-gold check. We additionally report an informative-gold check: gating base conditions by *acc_norm* (distinct from the *acc* endpoint, so conditions are not selected on the outcome), PRIORBPB's advantage is reliably signed on informative golds ($\Delta = +0.043$ on the standard Pythia ladders, $+0.032$ across all base families, $P(\Delta > 0) = 1.000$) and reliably larger there than on floor golds ($P = 0.96$–0.97).

Dense ladders resolve a saturation that 3-point ladders cannot. With three rungs, Spearman is ± 1 unless there is an exact tie, so the rank statistic cannot resolve small fidelity differences: 67% of the 3-rung ladders saturate $|-\text{Spearman}| = 1$. The ≥ 12-rung Pythia ladders do not (0%), which gives the test its power.

Result. On the primary gold (raw accuracy, $-$Spearman; Table 1, with the per-stratum breakdown in Table 3), the base-checkpoint comparison is reliably signed in PRIORBPB's favour: $\Delta = +0.021$, $P(\Delta > 0) = 1.000$, 95% credible interval $[+0.004, +0.079]$ ($-$Pearson agrees in sign). On instruct checkpoints, where a chat template injects a length-varying fixed cost, the advantage is larger ($\Delta = +0.122$, $P(\Delta > 0) = 1.000$, interval excluding zero on all three statistics), and the pooled posterior is likewise reliably signed ($\Delta = +0.047$, $[+0.011, +0.163]$, $P(\Delta > 0) = 1.000$). The base result is stable across the shared-prior mean $\mu \in \{40, 60, 111, 160, 220\}$ (all $P(\Delta > 0) \geq 0.99$).

Not a length artefact. We would withhold the claim if PRIORBPB looked better on raw accuracy but worse on the length-corrected or length-independent golds. It does not: on length-corrected Bayesian accuracy the base comparison is reliably positive ($\Delta = +0.019$, $P(\Delta > 0) = 0.979$), and on length-independent letter-label accuracy it is an exact tie ($\Delta = 0$), as it must be, since constant 2-byte golds make the two estimators monotone transforms, so PRIORBPB is never worse. The base advantage therefore survives length correction and is not a shared length artefact.

The advantage concentrates where the accuracy gold is informative. Base accuracy is an uneven signal: because the statistic is within-ladder rank co-movement it is level-invariant (base models merely scoring lower cannot shrink Δ), but a *noisier* gold can. Base letter-label accuracy is largely floor noise (75% of those ladders move by < 0.05; Spearman(acc, capability)$= +0.08$), whereas base cloze accuracy is informative ($+0.79$). A further check (App. K) uses an informativeness gate defined on *acc_norm*, distinct from the *acc* endpoint, so conditions are not selected on the outcome. On informative base conditions the comparison is reliably signed in PRIORBPB's favour ($\Delta = +0.043$ on the standard Pythia ladders, $+0.032$ across all base families, $P(\Delta > 0) = 1.000$), and the advantage is reliably larger on informative than on floor golds ($P(\Delta_{\text{info}} > \Delta_{\text{floor}}) = 0.96$–$0.97$).

How large is Δ on the metric's own scale? We deliberately avoid a practical-equivalence threshold, reading each stratum by its direction ($P(\Delta > 0)$) and its posterior mean and interval. To give the magnitudes a scale, the fidelity gap between the two incumbent estimators (mean-of-ratios BPB vs. the length-confounded corpus BPB) is only $+0.000$ (base) and -0.031 (instruct, i.e., corpus is if anything the better tracker). Against that reference, the base advantage ($+0.021$) already exceeds the entire spread among the standard estimators, and the instruct advantage ($+0.122$) is several times it, so on the metric's own scale both effects are non-negligible, the instruct one more so.

External validity: the accuracy gold is itself instruction-flavoured. One caveat limits how far this comparison generalises. The ground-truth signal here is downstream accuracy, but accuracy protocols embed an instruction-following task design, selecting or emitting the answer in a prescribed format (a letter, a cloze completion, an exact string). Base checkpoints are penalised on that format compliance beyond their raw next-token competence, so accuracy under-measures base capability and is a partly mis-specified gold for base models; this is the same mechanism that leaves base letter-label accuracy at floor above. The superiority question "which BPB variant best tracks accuracy?" is therefore most meaningful precisely for instruction-tuned models, where accuracy is a faithful capability signal, which is also the stratum where PRIORBPB wins clearly. For base checkpoints the verdict should be read as tracking of an instruction-flavoured gold rather than of pure capability, and the strong superiority claim is deliberately confined to the instruct stratum. Establishing a base-model advantage against a capability signal that is not itself an instruction-following task (e.g., a held-out perplexity or a task-completion metric with no format requirement) is left to future work.

A second, model-agnostic axis of superiority: length de-confounding. Proxy tracking is only one sense of 'better,' and the one just bounded to the instruct stratum. The other, established in Section 6 and independent of any accuracy gold, is construct validity: PRIORBPB is far less correlated with gold length than the estimators reported by evaluation frameworks. This axis does not inherit the instruction-following limitation because it never uses accuracy, so it speaks directly to base models, and it extends the three-scale Qwen3 length result to a multi-family set of base models. On the base checkpoints of this study (60 checkpoints across five families: Pythia, Qwen3-Base, SmolLM2, Qwen2.5, OLMo-2), the mean $|\text{Spearman}(\text{metric},\ \text{mean gold length})|$ over length-varying tasks is 0.68 for mean-of-ratios BPB and 0.76 for corpus BPB but only 0.18 for PRIORBPB; a paired family-clustered Bayesian bootstrap puts PRIORBPB strictly less length-confounded than mean-of-ratios BPB ($\Delta|\rho| = 0.25$, 95% CrI $[0.06, 0.42]$, $P = 0.996$) and than corpus BPB ($\Delta|\rho| = 0.26$, $[0.12, 0.45]$, $P = 1.000$). So even where accuracy is a poor signal for base capability, PRIORBPB is the sounder measurement. The two axes are complementary: a model-agnostic measurement-validity advantage (length de-confounding, on base and instruct alike) and a proxy-fidelity advantage (accuracy tracking, on instruct), rather than a single instruct-only claim.

Table 3: Paired tracking-fidelity advantage Δ of PRIORBPB over mean-of-ratios BPB by stratum (primary gold: raw accuracy, $-$Spearman). Δ is the posterior mean of fidelity(PRIORBPB) $-$ fidelity(mean-of-ratios) from a task-clustered Bayesian bootstrap, reported with its 95% credible interval and $P(\Delta > 0)$; $\Delta > 0$ favours PRIORBPB. n is the number of (model, task) conditions and "clust." the number of task clusters.

Stratum	n	clust.	Δ (post. mean)	95% CrI	$P(\Delta > 0)$
base	62	24	$+0.021$	$[+0.004, +0.079]$	>0.999
instruct	19	12	$+0.122$	$[+0.005, +0.571]$	1.000
pooled	81	24	$+0.047$	$[+0.011, +0.163]$	1.000

L WITHIN-TASK LENGTH STRATIFICATION

Our analysis and the reduction identity say the confound is a cross-task, different-length effect: corpus BPB inherits each task's own length mix through $\alpha/\bar{n}$. This appendix tests that as a controlled manipulation, with no new compute, by holding the task (hence the capability) fixed and changing only the gold-length distribution.

Design. For every (model, task) with a wide within-task length spread (ratio of the 90th to 10th percentile of gold content bytes ≥ 2.5; 17 tasks $\times$ up to 30 model checkpoints from the base, instruct and second-family roots, 183 conditions) we sort items by gold content length and split them into five equal-count strata. Capability is essentially constant across strata of one task, so a length-neutral per-byte score should be flat across strata, whereas corpus BPB contains the fixed cost $\alpha/\bar{n}$ (Eq. equation 1) and must fall as a stratum's mean length rises. We recompute corpus BPB, mean-of-ratios BPB and PRIORBPB (frozen geometric prior, μ=111) per stratum, and per (model, task) take the rank correlation ρ between each estimator and the stratum's mean length: $|\rho|$ near 0 is length-neutral, near 1 is length-driven. The paired estimand is $|\rho_{\text{corpus}}| - |\rho_{\text{PRIORBPB}}| > 0$ (Bayesian bootstrap, clustered by task).

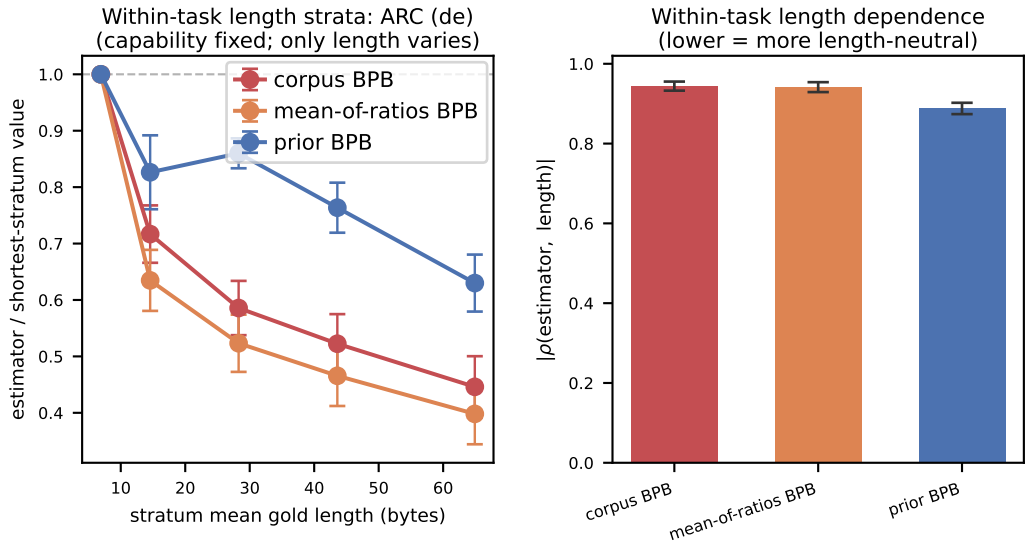


Figure 14: Within-task length stratification. **Left:** for the highest-dependence task, each estimator across five gold-length strata, normalised to its shortest-stratum value (capability fixed, only length varies); corpus BPB slides with stratum length while PRIORBPB is flatter. **Right:** $|\rho(\text{estimator}, \text{stratum length})|$ pooled over all 183 (model, task) conditions (lower is more length-neutral); PRIORBPB is the lowest.

Result. PRIORBPB is reliably the more length-neutral estimator within tasks: mean $|\rho|$ falls from 0.944 (corpus) and 0.941 (mean-of-ratios BPB) to 0.888 (PRIORBPB), a paired reduction of $+0.056$ (95% CrI $[0.015, 0.128]$, $P(\text{PRIORBPB more neutral})=0.99$; Fig. 14). The magnitude is modest by design: within one task the longer golds also differ in genuine content (a long GSM8K derivation is not a padded short one), so a per-byte *difficulty* score should retain some within-task length dependence, and what PRIORBPB removes is the $\alpha/\bar{n}$ fixed-cost slice of it. The large, clean effect is the cross-task decorrelation of Section 6 (mean $|\rho|$ from $0.68/0.76$ to 0.18), where the different-length confound of objection (iii) actually arises; the within-task stratum test here is the finer same-task control, and it agrees in direction. A complementary manipulation that instead prepends byte-matched prefixes to each gold is uninformative here: because $\mu=111$ far exceeds these gold lengths the geometric prior is near-flat over the relevant range, so PRIORBPB $\approx$ corpus BPB item-by-item and neither is invariant to padding, a reminder that a per-byte score is a within-task ranking device and does not provide a padding-proof capability score, reinforcing the recommendation to report the (α, β) split and an accuracy check alongside it.