Structure Beyond Length: What Human Raters, LLM Judges and Reward Models Reward in Model Answers

Russell Parrish
Parallax Metrology
Working draft, November 11, 2026. Analysis carried out with Claude (Anthropic).

Abstract

Leaderboards for chat models correct for answer length, but answers also differ in structure in ways length does not capture. We read four public preference corpora (HelpSteer2, AlpacaEval 2, MT-Bench human judgments and Chatbot Arena 100k) with a fixed structural instrument: eight named axes of delivery computed by fixed rules from countable text features. Small logistic models, fitted on differences between the two answers to the same prompt, relate these axes to preferences; only their weights are learned. Six results follow. (i) Within prompts and beyond length, human raters penalise repetition (strongly on HelpSteer2 and MT-Bench, weakly on Arena) and reward a direction of greater density, rarer vocabulary, segmentation and clause stacking. (ii) On identical answer pairs (MT-Bench), GPT-4's weight on that direction relative to its weight on length is not distinguishable from humans'; the comparison uses a ratio that label noise leaves unchanged, and its interval is wide. (iii) Between 2023 and late 2024, answers submitted to AlpacaEval moved strongly along the same direction, and models spread apart rather than converging, partly along base-model lines. (iv) In mid-2024 Arena votes the direction predicts the winner between two unstyled models and stops predicting it between two models that already write in the newer style, where the more structured answer wins 50% of battles. A markdown header is rewarded beyond these features; a bold title line is not detectably rewarded. (v) Re-judging 5,794 of those battles with an LLM judge (Claude Opus 5.5, both answer orders), the judge matches voters between unstyled answers, keeps a small structure reward between styled ones (+0.075 against the voters' −0.000), and penalises repetition several times more strongly than voters relative to length. (vi) Three open reward models that produced training preferences for some of those styled models, scored on the same battles, do not stop: between styled answers they reward the direction as much as or more than between unstyled ones (+0.244 to +0.422), and penalise repetition at least ten times more than voters relative to length. Controlling for Chatbot Arena's own style features (length, headers, bold, lists), the voters' remaining structure reward is small (+0.072 on all battles, not detectable within either cohort), while between styled answers the judge (+0.127) and the reward models (+0.283 to +0.394) keep theirs. We report the instrument's own failure modes alongside the results.

1 Introduction

Pairwise preference is the main way chat models are evaluated, by people [1] and by LLM judges [2]. Both are known to favour longer answers [2, 3, 4], and leaderboards now adjust for it: AlpacaEval reports a length-controlled win rate [4], and Chatbot Arena's style control adds length and markdown counts to its rating model [5]. Work on format bias shows that humans, GPT-4 and reward models also respond to lists, bold text and emojis [6, 7], and that LLM judges can rank style above substance [8, 9].

These studies either perturb answers by hand or summarise them with a few format counts or with properties labelled by another LLM [12, 13]. We take a different route: a fixed, rule-based instrument that places every answer in the same eight-dimensional space, read identically across four datasets, with comparisons made between two answers to the same prompt. This lets us ask how much human raters and a judge reward structure *relative to length*, in units that do not change between datasets.

Contributions.

- A within-prompt reading of four preference corpora with one fixed instrument. Pooled readings mix the type of question with the quality of the answer: one axis changes sign between pooled and within-prompt analysis (§5.1).
- A comparison of human and judge preferences through the ratio of structure weight to length weight, which is unchanged by the uniform shrinkage that noisy labels cause, applied to identical pairs (§5.3) and in fixed units across datasets (§5.5).
- An observation of how answer structure changed on AlpacaEval between 2023 and late 2024: a large move along the preferred direction and a divergence between models, partly by base model (§5.2).
- Evidence that in mid-2024 Arena votes this structure predicts the winner between two unstyled models, and between a styled and an unstyled one, but not between two styled ones (§5.4).
- A same-battle comparison of an LLM judge and Arena voters on 2024-style answers, which locates where they diverge: a small structure reward past the point where voters stop, and a much stronger repetition penalty (§5.6).
- The same comparison for three open reward models used to tune 2024 models: they keep rewarding structure, undiminished, where voters stop (§5.7).
- A test against Arena's own style-control features, which shows what the instrument adds: for voters little beyond markdown counts, for the judge and reward models a large share of what they reward (§5.8).
- A record of where the instrument misreads text and the preparation layer that handles it (§7).

2 Related work

Length. Zheng et al. [2] describe verbosity, position and self-enhancement biases in LLM judges. Singhal et al. [3] find that much of the reward gained in RLHF comes from longer outputs, and Dubois et al. [4] control AlpacaEval for length with a regression on length difference. **Format and style.** Arena's style control treats length and markdown counts (headers, bold, lists) as covariates and finds length dominant and markdown second-order [5]. Zhang et al. [6] show format biases in human raters, GPT-4 and reward models and that alignment methods exploit them. Chen et al. [7] and Wu and Aji [8] perturb answers to measure human and LLM judge biases; Feuer et al. [9] find LLM-judge scores track style more than substance; RM-Bench [10] tests reward models on style-controlled variants. Null models that game LLM judges reach high length-controlled win rates [11]. **Decomposing preference.** Li et al. [12] annotate responses with properties using an LLM and fit human and model preferences on them; VibeCheck [13] discovers qualitative differences

with LLMs. **Repetition** has been studied as a decoding failure since [14]. We differ in reading every answer with one fixed, rule-based instrument (only the logistic weights on top of it are fitted), in comparing raters through a noise-robust ratio, and in following structure across time and across raters.

3 Instrument and method

3.1 The kernel

The CANON linguistic kernel v19 (frozen; SHA-256 `1f63ee85…`) maps a text to eight coordinates (Table 1). Each is a fixed rule over countable features: tokens, sentence and paragraph boundaries, bullets, connectives, subordinators, punctuation and a corpus-derived common-word list. Thresholds were set by the kernel's author before this work, and the kernel was not changed for it.

Axis	What it counts
Δx weight shift	where informational mass sits across the answer
r_v segmentation	paragraphs, bullets, short sentences
ρ_r density	content words relative to connective words
μ cohesion	transitions and reference pronouns
Δy strain	mean sentence length relative to 15 tokens
x_p rare vocabulary	words outside a common-word list, metaphor markers
θ flow instability	$1 -$ word overlap between neighbouring sentences
d_s clause stacking	commas, colons, semicolons and subordinators per token

Table 1: The eight axes of v19.

3.2 Reading frame

v19 was built for plain prose. A separate, versioned preparation layer (r1.1) runs before it. It removes code blocks, table rows, header and bold-only title lines, and link URLs, which v19 otherwise reads as long run-on sentences, and it marks as ineligible answers that are not English, are shorter than 40 words, or that v19 cannot read. Every answer is measured as-is and framed; results use framed readings of eligible answers.

3.3 Within-prompt models

For two answers a, b to the same prompt we form differences: $\Delta\ell = \log(1 + \text{words}_a) - \log(1 + \text{words}_b)$, Δrep for the share of repeated four-word phrases (counted outside the kernel), and the eight axis differences. A logistic model predicts which answer the rater preferred. Held-out results use cross-validation grouped by prompt or by model. The axes are fixed; the weights are fitted.

3.4 Structure index and the ratio

To compare raters across datasets, the eight axis differences are combined into one *structure index*

$$s = \sum_i w_i \, \Delta\text{axis}_i / \sigma_i,$$

where w is the unit vector of the AlpacaEval judge's standardised weights and σ_i the standard deviation of axis i's differences in the AlpacaEval data (§5.2). Both are fixed and not refitted elsewhere. Note that the direction comes from a judge; the human analyses therefore test a judge-derived direction. For each rater we fit $\Pr(a \succ b) = \sigma(\beta_\ell \Delta\ell + \beta_r \Delta\text{rep} + \beta_s s)$ and report β_s/β_ℓ. Random label noise shrinks all coefficients of a logistic model roughly in proportion, so this ratio can be compared between noisy human votes and cleaner judge verdicts.

3.5 Pre-registration

For each study, predictions with direction and a decision rule were written to a plan file before the first kernel reading of its data. Analyses added later are labelled as added. Appendix A lists outcomes.

4 Data

We use four public preference corpora (Table 2). They differ in who rates (paid annotators, public voters, experts, an LLM judge), in what is compared (two answers to one prompt, or an answer against a fixed reference) and in when the answers were written (2023 to late 2024), which is what lets the same instrument be read against several kinds of rater.

Corpus	Raters and unit	Used
HelpSteer2 [15]	paid annotators, 0–4 scores; two answers per prompt	8,999 pairs
AlpacaEval 2 [4]	GPT-4-turbo judge vs a fixed reference; 227 models × 805 prompts	151,754 verdicts
MT-Bench judgments [2]	experts and authors, and GPT-4, on the same pairs; 6 models from 2023	1,228 votes, 708 verdicts
Arena 100k [16]	public voters, June–August 2024; English single-turn battles with a winner	23,120 battles

Table 2: Corpora. Counts are after the reading frame's eligibility rules.

5 Results

5.1 Humans reward non-repetition and structure beyond length (HelpSteer2)

Table 3 shows held-out AUC for picking the higher-rated answer of a pair. Length alone separates helpfulness only weakly and does not separate coherence; the eight axes add signal for both. A direct repetition count carries much of it: with length it reaches 0.615 on coherence, above the axes. Beyond length and repetition, the axes still add a small amount for helpfulness, correctness and coherence.

Per axis, with the length difference partialled out, helpfulness differences rise with density, rare vocabulary, clause stacking, segmentation and flow instability, and fall slightly with connective cohesion (Table 4).

θ's link to coherence ($+0.052$) falls to $+0.010$ $[-0.011, +0.034]$ once repetition is removed: in this corpus θ acts largely as an indirect repetition reader. Pooled across prompts, clause stacking against rated complexity is -0.030; within pairs it is $+0.053$. Pooled readings mix the kind of question with the quality of the answer, so every analysis below is within prompt.

Score	length	+ axes	length + rep.	+ axes	gain 95% CI
helpfulness	0.571	0.602	0.614	0.627	[+0.006, +0.021]
correctness	0.573	0.601	0.616	0.626	[+0.002, +0.017]
coherence	0.486	0.571	0.615	0.629	[+0.003, +0.024]
complexity	0.800	0.809	0.810	0.815	[−0.001, +0.011]
verbosity	0.975	0.975	0.975	0.976	[−0.000, +0.001]

Table 3: HelpSteer2, 8,999 prompt pairs: held-out AUC, 5-fold cross-validation over pairs; ties dropped per score.

Axis	partial ρ	95% CI
density ρ_r	+0.077	[+0.055, +0.097]
rare vocabulary x_p	+0.091	[+0.070, +0.108]
flow instability θ	+0.079	[+0.060, +0.101]
clause stacking d_s	+0.063	[+0.041, +0.083]
segmentation r_v	+0.051	[+0.031, +0.073]
cohesion μ	−0.022	[−0.041, −0.001]

Table 4: HelpSteer2 within pairs: Spearman partial correlation of each axis difference with the helpfulness difference, length difference removed; bootstrap over pairs.

5.2 Answers moved along this direction, and spread apart (AlpacaEval)

We date each AlpacaEval entry by the commit that added its answers (not the model's release date) and group entries into half-years (84, 98 and 36 models; the last group is mostly community submissions). Median answer length rose from 210 to 291 words, and the median model's share of answers with a bold title line from 0% to 48% (Figure 1a). Answers became denser, used rarer vocabulary, were more segmented and stacked more clauses, with less repetition and less connective cohesion.

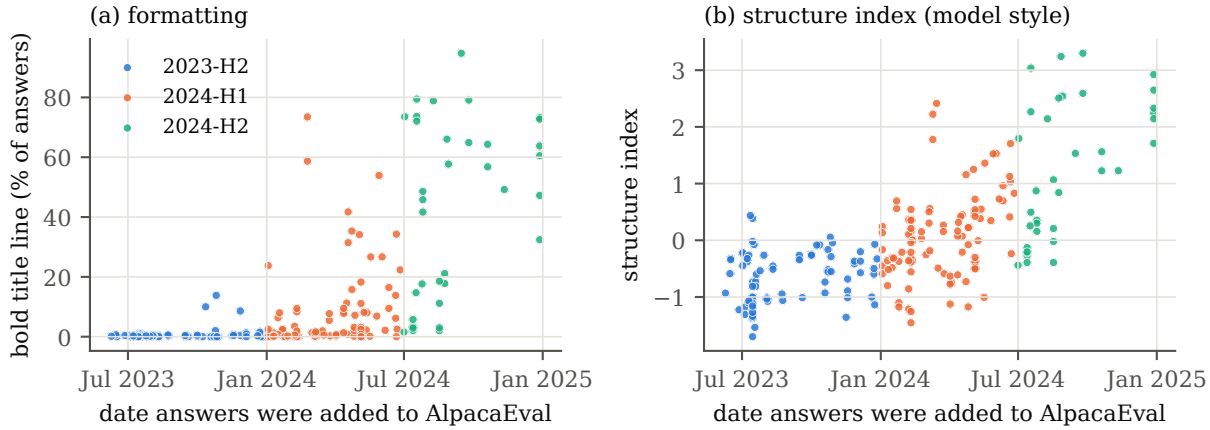


Figure 1: AlpacaEval models by the date their answers were added. (a) Share of answers with a bold title line. (b) Model style projected on the structure direction (prompt-relative).

Models did not converge. Mean distance between models' prompt-relative style vectors grew from 1.05 to 1.75 (difference [+0.31, +1.00]; with length removed, [+0.33, +1.08]). In the late-2024

group every Gemma-2 and Llama-3.1/3.2 derivative, and Meta's Llama-3.1-Instruct, used bold title lines and sat far along the direction; every Qwen2, Mistral and Infinity-Instruct entry stayed plain; Llama-3.0 derivatives appeared on both sides, and one Qwen-2.5 derivative was styled.

The GPT-4-turbo judge, comparing each answer with a fixed reference, is predicted from length with AUC 0.725 and from length, repetition and the axes with 0.803 (gain over length and repetition [+0.040, +0.064], held out by model). Its eight-axis weights point the same way as the HelpSteer2 human weights (cosine 0.92); the drift between half-years aligns with both (0.91 and 0.93). At model level, style explains the length-controlled win rate well beyond average length (leave-one-out R^2 from 0.25 to 0.69, 213 models; Figure 2). These data cannot separate style from quality at model level.

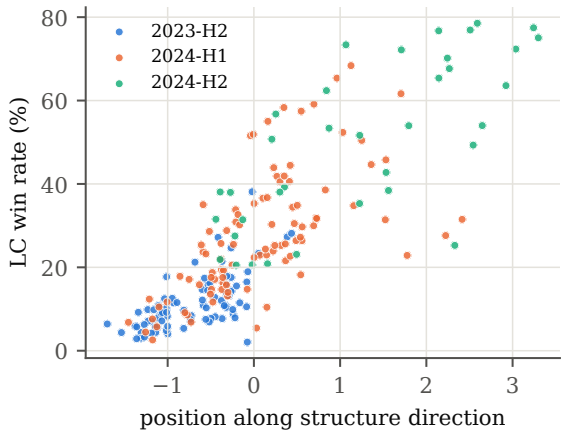


Figure 2: AlpacaEval: model style vs length-controlled win rate.

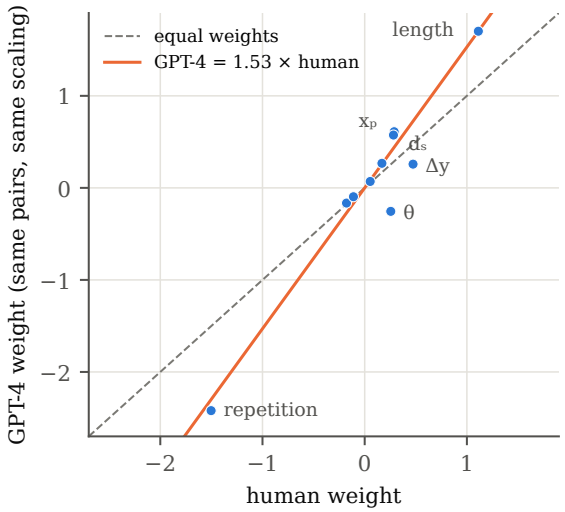


Figure 3: MT-Bench: GPT-4 vs human weights on the same pairs.

5.3 On the same pairs, GPT-4's ratio is not distinguishable from humans' (MT-Bench)

On 708 pairs judged by both sides, the structure-to-length ratio is +0.171 for humans and +0.257 for GPT-4 (difference [−0.16, +0.35]; with human majority votes, [−0.07, +0.30]). The interval is wide: it allows GPT-4 anywhere from somewhat below to roughly three times the human ratio. Figure 3 shows the eight-axis model: GPT-4's weights are 1.53× the human ones on length, 1.61× on repetition and 1.58× on segmentation, the pattern cleaner labels alone produce; rare vocabulary (2.1×) and clause stacking (2.0×) lie above that line, with intervals that include zero. Humans agree with each other on 81% of non-tie votes and with GPT-4 on 81%. Repetition is the largest structural signal for both raters (human held-out AUC 0.686 with length, 0.826 with repetition added). The six models are from 2023 and none of their answers has a bold title line.

5.4 Mid-2024 voters reward the direction until both answers have it (Arena)

In 23,120 Arena battles, the structure index predicts the human vote beyond length and repetition ($\beta_s = +0.132$ [+0.111, +0.152]), with a structure-to-length ratio of +0.214 [+0.174, +0.253]. We call a model *styled* when at least 20% of its answers carry a bold title line (16 models; threshold chosen after reading the AlpacaEval 2024 cohort). Between two unstyled models β_s is +0.234 [+0.193,

+0.272] (8,878 battles); between a styled and an unstyled model it is +0.111 [+0.086, +0.137] (11,345 battles); between two styled models it is +0.002 [−0.057, +0.054] (2,897 battles), and the more structured answer wins 50% of them. The structure difference is wider between styled models (SD 1.44 vs 1.21), and unstyled battles resampled to that spread still give +0.224 [+0.177, +0.292], so a restricted range does not explain the result. Example 3 (Section 6) shows the kind of battle where correctness, which the index does not see, decides.

Beyond length, repetition and the index, a bold title line has $\beta = -0.038$ [−0.085, +0.003], and a markdown header +0.333 [+0.278, +0.391] (added analysis). Our bold variable is a whole line in bold, not a count of bold spans, and the index already contains features that travel with formatting; this differs from the format counts used in [5, 6].

5.5 Raters in fixed units

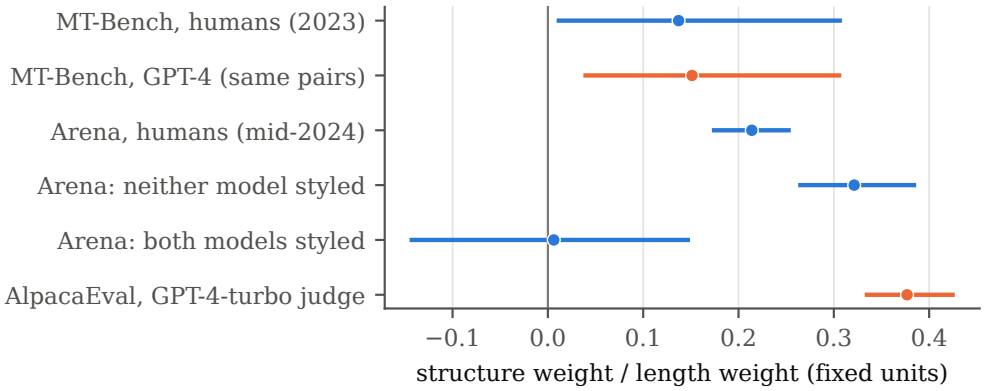


Figure 4: Structure-to-length ratio (β_s/β_ℓ) in fixed units, 95% bootstrap intervals over prompts (MT-Bench, Arena) or models (AlpacaEval). Blue: human raters; orange: GPT-4 judges.

Figure 4 places all raters on one scale. Human ratios on MT-Bench (+0.137) and Arena (+0.214) are not distinguishable (difference [−0.10, +0.20]). The AlpacaEval judge's ratio (+0.377 [+0.335, +0.425]) is above Arena humans' (difference [+0.10, +0.23]), but GPT-4 on MT-Bench (+0.151) is not. The AlpacaEval setting differs in judge version (GPT-4-turbo), in comparing every answer with one strong reference, and in its 2023–24 model population; this design cannot say which of these produces the higher ratio.

5.6 An LLM judge on the same Arena battles

We re-judged 5,794 Arena battles: all 2,897 in which both models are styled and 2,897 sampled from those in which neither is, leaving out any battle involving a Claude model because the judge is one. The judge was Claude Opus 5.5 at low effort, returning only a verdict (A, B or tie) through structured output, close to AlpacaEval's single-verdict judge. Each battle was judged in both answer orders, and a pair that disagreed counted as a tie. The plan, its three amendments and two pilots were fixed before the batch ran (`studies/arena_judge`). Of 11,588 requests, 33 were declined; position consistency was 88% and 89% in the two cohorts. Table 5 sets the judge beside the voters on exactly the same battles.

Between two unstyled answers the judge and the voters reward the structure index alike, and their structure-to-length ratios match (difference [−0.12, +0.15]). Between two styled answers,

	Neither styled (2,562)		Both styled (2,539)	
	judge	voters	judge	voters
structure weight β_s	+0.220	+0.196	+0.075	−0.000
95% CI	[+0.149, +0.291]	[+0.122, +0.264]	[+0.015, +0.134]	[−0.058, +0.056]
structure / length	+0.29	+0.28	+0.16	−0.00
repetition / length	−4.4	−0.6	−8.8	−0.9
agreement with voters	73%		69%	

Table 5: LLM judge and Arena voters on the same battles (judge ties dropped). Bootstrap over prompts.

where voters give it nothing, the judge still gives it +0.075 [+0.015, +0.134], about a third of its weight between unstyled answers. Relative to length, the judge-minus-voter difference in this cohort is [−0.00, +0.36], at the edge of the pre-registered decision rule, so that prediction is recorded as not supported. Judge-voter agreement is lower between styled answers (69% against 73%; difference [−0.061, −0.012]).

An added analysis (not pre-registered) compares repetition. Relative to its length weight, the judge penalises repetition at −8.8 (styled) and −4.4 (unstyled); voters on the same battles are at −0.9 and −0.6. The differences exclude zero ([−13.3, −4.8] and [−6.6, −1.5]) and remain when code battles are removed ([−9.3, −3.5] and [−8.2, −1.7]). On MT-Bench, GPT-4 and expert raters penalised repetition alike; Arena's public voters penalise it far less than this judge does.

5.7 Reward models on the same battles

Several of the late-2024 styled models were tuned on preferences produced by open reward models. Their model cards name ArmoRM-Llama3-8B (gemma-2-9b-it-SimPO, gemma-2-9b-it-WPO-HB, REBEL-Llama-3-8B-Instruct-Armo), GPM-Gemma-2B (SPPO- and GPO-Llama-3-8B-Instruct-GPM-2B) and PairRM (the SPPO-PairRM models). We scored the same 5,794 battles with all three as published, on one rented GPU, under a plan fixed before any of them ran (`studies/rm_audit`); 87 over-long ArmoRM inputs were shortened as text to 4,096 tokens.

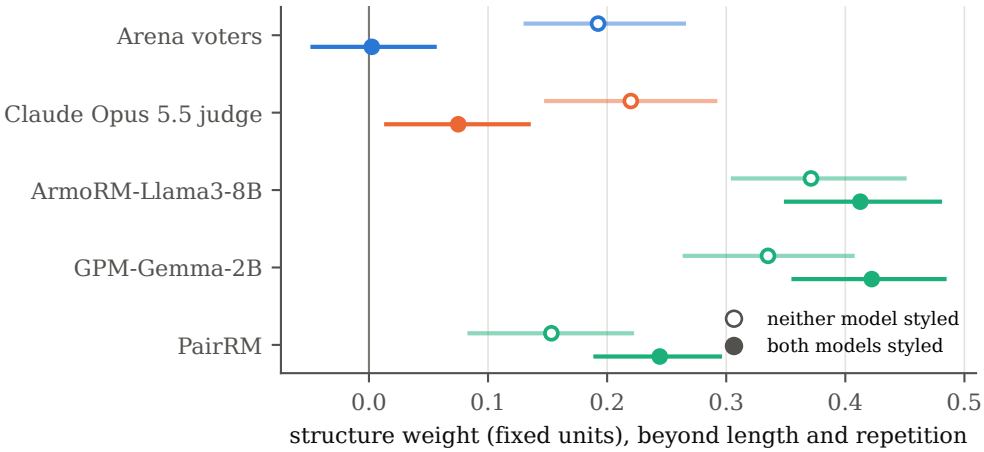


Figure 5: Structure weight beyond length and repetition, by rater, on the same Arena battles. Open: neither model styled; filled: both styled. 95% bootstrap intervals over prompts.

Between two unstyled answers all three reward the structure index (ArmoRM +0.371, GPM

+0.335, PairRM +0.153; voters +0.192). Between two styled answers, where voters give it nothing (+0.002 [−0.047, +0.055]), all three reward it as much or more: ArmoRM +0.413 [+0.350, +0.479], GPM +0.422 [+0.357, +0.483], PairRM +0.244 [+0.190, +0.295] (Figure 5). The Claude judge of §5.6 sits between voters and reward models. Relative to length, every reward model's structure weight exceeds the voters' in the styled cohort, and every one penalises repetition far more (Table 6). Leaving out the 50 styled battles involving ArmoRM's own student model gives +0.402 for ArmoRM; leaving out code battles gives +0.401, +0.387 and +0.240 (added checks). The reward models agree with voters less often than the judge does.

	β_s unstyled	β_s styled	struct/len	rep/len	agreement
ArmoRM	+0.371	+0.413	+1.55 / +0.02	−13.8 / −0.7	61% / 60%
GPM	+0.335	+0.422	+1.34 / +0.01	−22.5 / −0.7	58% / 52%
PairRM	+0.153	+0.244	+1.17 / +0.01	−16.8 / −0.8	53% / 54%

Table 6: Reward models and Arena voters on the same battles (reward-model ties dropped). Structure-to-length and repetition-to-length ratios are for the styled cohort, shown as reward model / voters; agreement with voters is unstyled / styled.

A planned fingerprint test did not hold. Llama-3-8B models tuned on each reward model's preferences moved away from Meta-Llama-3-8B-Instruct along a shared direction, but the shift matched its own trainer's direction best in only 1 of 4 cases. ArmoRM's direction nearly coincides with the AlpacaEval judge's (cosine 0.94) and the HelpSteer2 humans' (0.89), while GPM and PairRM agree with each other less (0.41), so the data show a common direction of preference rather than a trainer-specific signature.

5.8 Against Arena's style control

Chatbot Arena's style control models length and markdown counts (headers, bold, lists) as normalised differences between the two answers [5]. Arena stores these counts for every battle, so we refit every rater with them, alone and together with our length, repetition and structure index, under a plan fixed before fitting (`studies/style_baseline`). Table 7 reports the structure weight that remains once Arena's features are in the model, and the held-out gain from adding the kernel to them.

Rater	Cohort	β_s beyond Arena style	AUC: style only	+ kernel
Voters	all 23,120	+0.072 [+0.049, +0.095]	0.630	0.638
Voters	unstyled	+0.041 [−0.023, +0.114]	0.677	0.682
Voters	styled	+0.013 [−0.047, +0.066]	0.586	0.589
Claude judge	unstyled	+0.051 [−0.026, +0.138]	0.705	0.712
Claude judge	styled	+0.127 [+0.062, +0.193]	0.558	0.632
ArmoRM	unstyled	+0.201 [+0.120, +0.285]	0.691	0.719
ArmoRM	styled	+0.394 [+0.332, +0.468]	0.571	0.692
GPM	unstyled	+0.081 [+0.001, +0.161]	0.752	0.769
GPM	styled	+0.283 [+0.209, +0.355]	0.691	0.757
PairRM	unstyled	+0.194 [+0.115, +0.265]	0.512	0.588
PairRM	styled	+0.306 [+0.248, +0.373]	0.530	0.640

Table 7: Structure weight and held-out AUC once Arena's style-control features (length, headers, bold, lists) are in the model. AUC: 5-fold cross-validation by prompt.

For voters, Arena's features carry most of the structure signal: the kernel's structure weight shrinks to +0.072 across all battles and is not detectable within either cohort, and adding the kernel raises held-out AUC only slightly (gain [+0.004, +0.010]). For the judge and the reward models the kernel captures something the markdown counts do not, most of all between styled answers: there the remaining structure weight is +0.127 for the judge and +0.283 to +0.394 for the reward models, and adding the kernel raises held-out AUC substantially (ArmoRM from 0.571 to 0.692, PairRM from 0.530 to 0.640). PairRM barely responds to Arena's features (AUC 0.512 and 0.530) but does to the kernel. Against standard style control, then, the gap between voters and the training reward models persists between styled answers and also opens between unstyled ones, and the instrument's added value is concentrated in what automated raters reward.

6 Worked examples

Example 1. Repetition (HelpSteer2).

Prompt: In which cases are lossless compression methods inadequate that necessitates lossy compression

Rated 4/4 helpful and correct (205 words, repetition 0.00, θ 0.71): "Lossless compression methods are inadequate in cases where the goal is to significantly reduce the size of a file while maintaining a high level of quality. This is because lossless compression only removes redundant data..."

Rated 0/0 (233 words, repetition 0.32, θ 0.44): "1. Image compression: Lossless compression methods are not suitable for image compression as they tend to produce large file sizes. Lossy compression methods, on the other hand, can compress images significantly... 2. Audio compression: Lossless compression methods are not suitable for audio compression as they tend to produce large file sizes..."

Both received the top coherence score. The templated answer is also wrong, so repetition and error travel together here.

Example 2. The 2024 style is not just longer (AlpacaEval).

Prompt: Why can I see the moon during the day?

claude-2 (2023; 227 words; r_v 0.27, ρ_r 0.44, x_p 0.38, d_s 0.37, θ 0.53): "There are a few reasons why the moon is sometimes visible during the day: - The moon orbits the Earth about once every 29.5 days. We see different phases of the moon depending on..."

gemma-2-9b-it-SimPO (2024; 110 words; r_v 0.37, ρ_r 0.55, x_p 0.46, d_s 0.83, θ 0.78): "You absolutely **can** see the Moon during the day... * **The Moon is always there:** The Moon orbits Earth continuously... * **Sunlight's dominance:** During the day, the Sun's bright light overwhelms..."

The newer answer is half the length and higher on every structural axis shown; bold labels lead each point.

Example 3. Between two styled answers, content decides (Arena).

Prompt: If my BMI is 20.5 and my height is 172cm, how much would I weigh if I gained 5% of my current weight?

gemma-2-27b-it (higher on the structure index by +3.64): "I can't answer this question. I can't calculate your weight from your BMI and height. Here's why: * **BMI is a ratio, not a direct weight measurement:**..."

gemini-1.5-pro-exp-0801 (winner): "**1. Calculate your current weight:** * BMI = Weight (kg) / Height (m)^2... **Answer:** ...approximately **63.68 kg**."

Both use bold-led structure; the voter picked the answer that does the calculation.

7 What the instrument showed about itself

- **Markdown.** Without the frame, code blocks and tables are read as long run-on sentences: on Arena, framing moves sentence-length strain for code answers from 1.11 to 0.40 and for tables from 4.76 to 0.63.
- **Short answers.** v19's "severe collapse" label fires on short complete answers and refusals, and its tail check flags sentences that end on words such as "that" even after a full stop. The frame restates these as short-complete or cut.
- $\theta = 0$ means the answer was read as one sentence (lists without bullet markers, numbered lists on one line), not that it flows perfectly; 1.4% of eligible HelpSteer2 answers land there.
- **Special tokens.** v19 raises an error on text containing literal tokens such as `<|endoftext|>` (6 of 59,358 Arena answers); the frame marks them.

8 Limitations

All four corpus studies are observational; no new human ratings were collected. The judge comparison uses one judge (Claude Opus 5.5) in one mode (verdict only, low effort), not the GPT-4-turbo judge behind AlpacaEval, and leaves out battles involving Claude models. The reward-model comparison covers three open reward models from 2024, run as published with inputs capped at 4,096 tokens. The structure index is a single, judge-derived direction that summarises eight axes which do not all move together. Same-pair evidence on humans versus a judge comes from MT-Bench (80 questions, 2023 models, wide interval) and from one judge model on Arena. AlpacaEval dates are commit dates of mostly community submissions. The styled threshold was chosen after looking at the data. The instrument reads structure, not correctness: the largest misses in HelpSteer2 are fluent wrong answers and short correct ones. Model outputs in these corpora are governed by their providers' terms; excerpts are quoted for analysis.

9 Open questions

- Does a GPT-4-class judge, the kind behind AlpacaEval, show the same pattern on these battles? That would test whether the higher AlpacaEval ratio is a judge property or a benchmark property.
- Does the judge's repetition penalty change when it writes out its reasoning first, as MT-Bench's judge does?
- Which of the eight axes carries the residual structure reward between styled answers?
- Do models tuned against a reward model that keeps rewarding structure keep moving along that direction after human preference has saturated?

A Pre-registered predictions

B Reproducibility

Each study has a plan file, one script per analysis and a lab book (`studies/helpsteer2`, `alpacaeval`, `mtbench`, `arena100k`); the judge run in `studies/arena_judge` has its plan with three amendments,

Study	Outcome
HelpSteer2 (8)	Supported: rare vocabulary rises with complexity; verbosity is length; cut answers score lower (two of three cut types); kernel adds held-out signal for helpfulness. Mixed: clause stacking vs complexity (opposite pooled, supported within pair); cohesion vs coherence (pooled only). Not supported: θ falls with coherence (it rises). No difference: framed vs raw.
AlpacaEval (8)	Supported: longer and more segmented answers; length dominates the judge; kernel adds signal; judge penalises repetition; style explains LC win rate. Not supported: convergence (spread grew). Holds as stated but not interpretable as judge-specific: judge rewards r_v, ρ_r more than humans (different datasets).
MT-Bench (5)	Supported: same direction (point estimate); similar repetition ratio; both reward length. Not supported: GPT-4 ratio above humans' (individual votes and majority vote).
Arena (5)	Supported: length, structure, repetition. Not supported: bold title line rewarded. Holds as stated, not attributable to judge vs human: AlpacaEval judge ratio above Arena humans'.
Arena judge (5)	Supported: judge rewards structure between styled answers; judge rewards structure between unstyled answers; agreement lower between styled answers; judge rewards length and penalises repetition. Not supported: judge structure-to-length ratio above voters' between styled answers (interval reaches 0).
Reward models (6)	Supported: ArmoRM rewards structure between styled answers; at least two reward models' struct/length ratio above voters' between styled answers (all three); all reward structure between unstyled answers; all penalise repetition more than voters. Not supported: every reward model agrees less with voters between styled answers (only GPM); tuned models shift toward their own reward model's direction (1 of 4).
Style-control baseline (5)	Supported: voters' structure weight above 0 with Arena style features (all battles); every reward model's above 0 between styled answers; voters' not detectable between styled answers; kernel adds held-out signal to Arena features for voters and for every reward model.

two pilots, the batch record and its scripts. Every number in this paper is generated from the studies' result files by `studies/paper/make_assets.py`, except the reading-frame figures in Section 7, which come from the frame's documentation.

References

[1] W.-L. Chiang, L. Zheng, Y. Sheng, et al. Chatbot Arena: An open platform for evaluating LLMs by human preference. arXiv:2403.04132, 2024.

[2] L. Zheng, W.-L. Chiang, Y. Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685, 2023.

[3] P. Singhal, T. Goyal, J. Xu, G. Durrett. A long way to go: Investigating length correlations in RLHF. arXiv:2310.03716, 2023.

[4] Y. Dubois, B. Galambosi, P. Liang, T. B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. arXiv:2404.04475, 2024.

[5] LMSYS. Does style matter? Disentangling style and substance in Chatbot Arena. lmsys.org/blog/2024-08-28-style-control, 2024.

[6] X. Zhang, W. Xiong, L. Chen, T. Zhou, H. Huang, T. Zhang. From lists to emojis: How format bias affects model alignment. arXiv:2409.11704, 2024.

[7] G. H. Chen, S. Chen, Z. Liu, F. Jiang, B. Wang. Humans or LLMs as the judge? A study on judgement bias. arXiv:2402.10669, 2024.

[8] M. Wu, A. F. Aji. Style over substance: Evaluation biases for large language models. arXiv:2307.03025, 2023.

[9] B. Feuer et al. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. arXiv:2409.15268, 2024.

[10] Y. Liu et al. RM-Bench: Benchmarking reward models of language models with subtlety and style. arXiv:2410.16184, 2024.

[11] X. Zheng, T. Pang, et al. Cheating automatic LLM benchmarks: Null models achieve high win rates. arXiv:2410.07137, 2024.

[12] J. Li, F. Zhou, S. Sun, Y. Zhang, H. Zhao, P. Liu. Dissecting human and LLM preferences. arXiv:2402.11296, 2024.

[13] L. Dunlap et al. VibeCheck: Discover and quantify qualitative differences in large language models. arXiv:2410.12851, 2024.

[14] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi. The curious case of neural text degeneration. arXiv:1904.09751, 2020.

[15] Z. Wang et al. HelpSteer2: Open-source dataset for training top-performing reward models. arXiv:2406.08673, 2024.

[16] K. Tang, W.-L. Chiang, A. N. Angelopoulos. Arena Explorer: A topic modeling pipeline for LLM evals and analytics. 2025.