Dentio

REMOVING BIAS BY ADDING BIAS

Style normalization for less noisy LLM-as-a-judge evals

Fabian Farestam, Jonathan Ahrlind

Dentio

Correspondence: fabian@dentio.com

Date: May 6, 2026

ABSTRACT

LLM-as-a-judge evaluation has become a default pattern for production AI systems but is susceptible to style bias, often rewarding verbosity or favoring certain model families. This poses a challenge when evaluating clinical documentation workflows, such as those at Dentio, where various writing styles for clinically correct journal notes can easily drown out the signal of factual preservation. Our solution is style normalization: rewriting each candidate into short, neutral factual prose before passing it to the judge. Instead of aiming for an unbiased judge, we introduce a controlled bias that is applied uniformly across all candidates, which is more useful for production evaluations driven by relative comparisons. Testing on an internal set of manually annotated Swedish dental journals showed that without normalization, the same factual content received scores spread across 2.5 points on a 10-point scale depending purely on style. Normalization reduced this style-induced score spread by up to 53% using larger normalizer models. Crucially, this process reduced style bias without reducing the judge's sensitivity to clinically meaningful factual errors.

1 Introduction

LLM-as-a-judge evaluation has become a default pattern for many production AI systems. It is useful, cheap enough to run often, and flexible enough to score messy outputs that do not fit clean unit tests. It is also easy to fool, as judges tend to prefer certain styles, often rewarding verbosity, and can favor outputs that look like something from their own model family [1-7]. In multilingual settings, including Swedish, the agreement can get worse [8, 9].

At Dentio, this matters because we are evaluating clinical documentation workflows, not chatbot answers. We help dentists draft journal notes from patient conversations, and a clinically correct note can look very different depending on the dentist. It might be terse or verbose, bullet-pointed or narrative, structured by procedure or by chronology. This variation, while good for the product, is bad for evals. If the goal is to measure whether the pipeline preserves the right clinical facts, then style should not influence the score as much as it usually does. Our fix is deliberately simple: before judging, make every candidate boring.

We call this *style normalization*: rewriting each candidate into short, neutral factual prose before passing it to the judge. Instead of pretending the judge is unbiased, we introduce a controlled bias that is applied uniformly across all candidates. For production evaluations, where model selection is often driven by relative comparisons, this method can be more useful than an ostensibly neutral judge that secretly prefers one writing style.

2 The eval problem we ran into

Some evaluations can be made very crisp, such as factual checks like whether the note mentioned the correct tooth number, included the diagnosis, or captured the planned treatment.

However, many production evaluations are messier. A good dental journal note needs the right clinical content, structure, level of specificity, and amount of uncertainty to be useful to the dentist who reviews it later; it is not just a bag of facts. This creates an annoying evaluation problem: style is sometimes part of quality, and sometimes it is just noise.

When evaluating final user-facing notes, style matters. But when evaluating whether a pipeline preserves clinical information across model changes, style can easily drown out the signal, causing a more fluent or verbose candidate to score higher even if the underlying clinical content is identical.

Researchers have tackled this from several directions: checklist-style evaluators force the judge to score smaller pieces [10]; pairwise comparisons often align better with human judgment, though they introduce ordering and position effects [11]; calibration, reference answers, multi-agent judging, and repeated sampling can reduce variance and make scores more stable [12, 13]; atomic factuality methods highlight how easily holistic scoring can mix factual

accuracy with fluency [14]; and fine-tuned judges can help with cost and latency but can also learn their own domain-specific biases [15].

These methods are useful, but dental documentation is rich enough that reducing everything to canonical claims can lose nuance, while broad holistic scoring allows style bias to creep back in. We sought a smaller layer to put before an existing judge to make style less visible when style was not the factor being tested.

3 Removing / Adding bias

The setup is simple. To evaluate some factor A (e.g., whether a generated journal note preserves clinically relevant facts), instead of sending the candidate directly to the judge, we first send it to a separate model Y with a narrower instruction:

> *Du är en medicinsk textnormaliserare. Din uppgift är att läsa ett journalanteckningsutdrag och återge ENBART de kliniska fakta som en kompakt, neutral punktlista på svenska.*
>
> *Regler:*
> *- En rad per faktum. Använd bindestreck som punkt.*
> *- Behåll tandnummer, ytor, diagnoser, mätvärden och föreslagna åtgärder exakt som de står.*
> *- Ta bort artighetsfraser, motiveringsspråk, stilistiska omsvep och formateringsrubriker som "Anamnes" / "Bedömning".*
> *- Lägg INTE till information som inte finns i originalet.*
> *- Skriv inga förklaringar eller rubriker före listan. Bara raderna.*
>
> *Journaltext:*
> *---*
> *{journal}*
> *---*

The output from Y is then passed to the same LLM-as-a-judge evaluator.

The normalizer is not supposed to improve the answer; its job is narrower and more boring: preserve the factual content while removing presentation differences that should not matter for this evaluation. This still introduces bias, as the normalized text will reflect the preferences and failure modes of model Y, but this bias is applied uniformly across all candidates.

For production evaluations, especially regression tests and A/B comparisons, stability of score differences is often more critical than the absolute calibration of the score. A normalizer that forces every candidate through the same style bottleneck gives us a bias that is easier to reason about.

This approach also reduces self-preference bias. Without normalization, a judge that prefers outputs that look like its own model family will have that preference leak into the score. With normalization, every candidate is first rewritten by the same model Y, ensuring that any "Y-style" bias is applied to everyone. After normalization, this remains a regular judge problem, and techniques like decomposing criteria, using references, or running pairwise comparisons can still be applied [13]. The normalization layer merely makes one common source of noise less visible.

4 Some results

We tested this on an internal set of manually annotated Swedish dental journals. For each example, we generated style variants (verbose, terse, bullet, narrative) holding all clinical facts identical, then scored every variant against the gold reference. This resulted in four scored variants per journal for each normalizer.

The judge used a 1-10 holistic clinical-quality score covering correctness, structure, and professionalism. This score was intentionally not a perfectly isolated factuality score, resembling the kind of messy production evaluations people actually use. The fact that style normalization still reduced variance under this holistic judge makes the result more practically relevant.

4.1 The baseline problem

Without normalization, the same factual content received scores spread across **2.5 points** on a 10-point scale depending purely on style. This 25% noise from presentation alone is enough to make A/B comparisons between pipeline versions unreliable.

4.2 Normalization reduces style bias

We tested ten different models as the normalizer Y, ranging from small (Gemma 3 4b, GPT-5.4 nano) to large (Claude Opus 4.7, GPT-5.5). In this sample, every tested model reduced style-induced spread, with larger normalizers generally performing better. Spread reduction is measured as 1 minus the ratio of score range after normalization to score range before normalization, so 53% means the style-induced scoring gap shrank by about half.

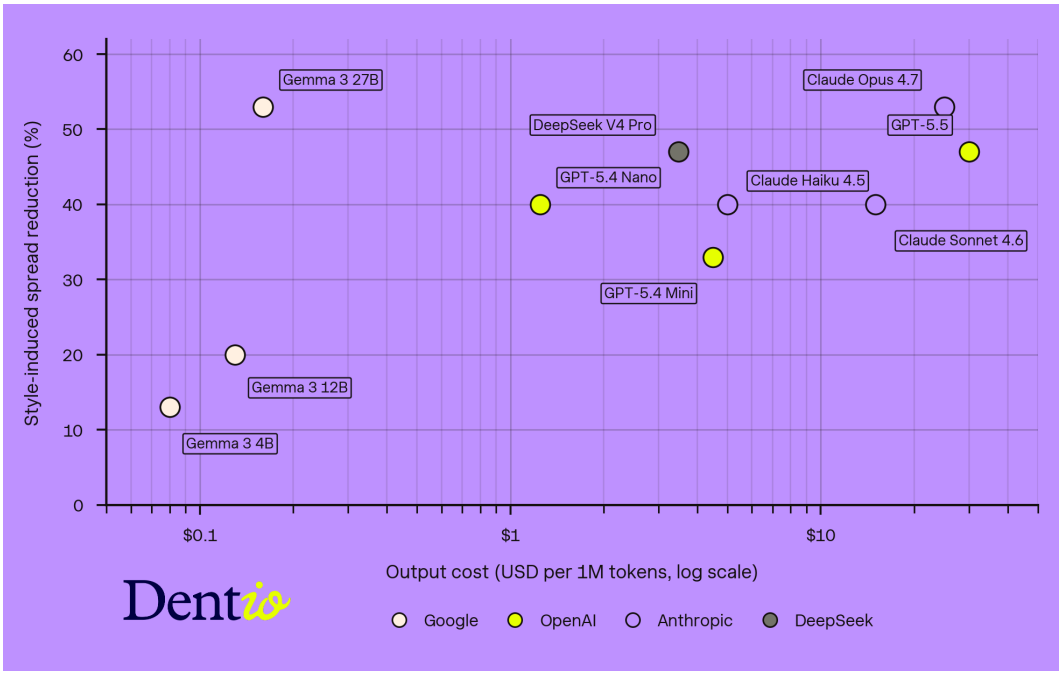

Figure 1

Spread Reduction by Normalizer Model Size

The pattern is clear: larger normalizers are better at stripping style without distorting content, but even a small model provides meaningful improvement.

4.3 Factual sensitivity is preserved

The critical safety question is whether normalization buries real clinical errors. We tested this by introducing wrong-tooth errors into the journals (changing one FDI tooth number to another valid but incorrect number) and measuring how much the judge's score dropped compared to the clean version. We then compared that score drop with normalization against the score drop without normalization. A value of 100% means the judge catches the error equally well in both conditions.

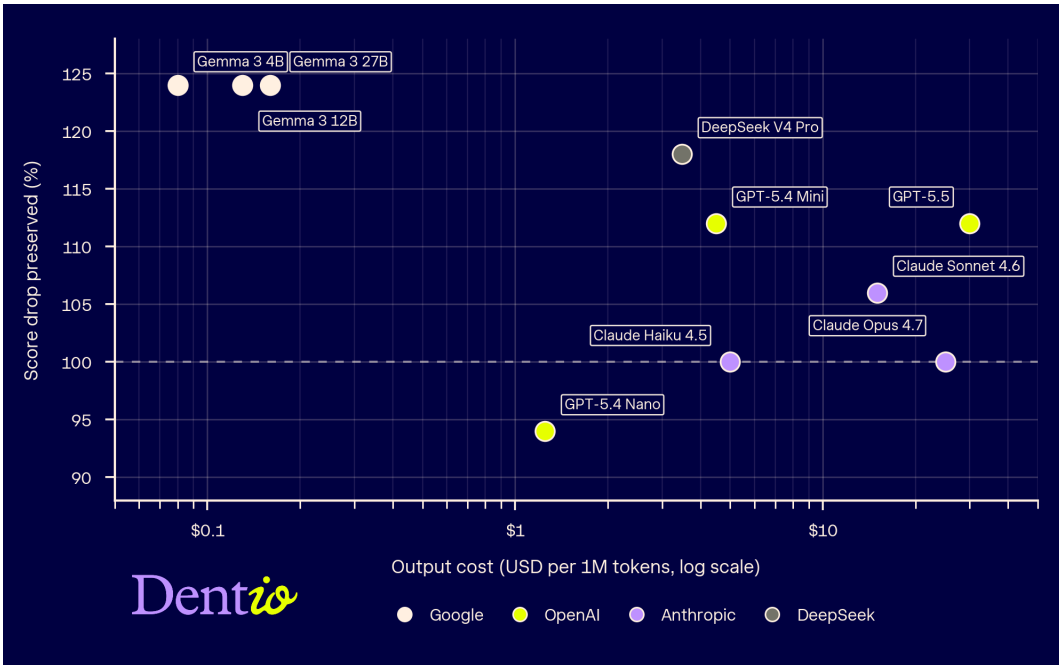

Figure 2

Error Preservation by Normalizer Model Size

Values above 100% mean the score drop for the wrong-tooth version was larger after normalization than before, suggesting the factual error became more salient once stylistic variation was stripped away. The important result is that normalization reduced style-induced score spread without reducing sensitivity to a clinically meaningful injected error.

4.4 Practical takeaway

You do not need a frontier model as the normalizer. In this setup, mid-size models gave 33-40% spread reduction while preserving sensitivity to the error class we tested, at a fraction of the cost of using models like Claude Opus 4.7 or GPT-5.5 for every eval call.

5 Conclusion

Style normalization is one small layer, not the whole eval stack. But it helped with a very real production problem: our evals were too easily impressed by style. By making every candidate boring first, we made the judge focus more on the clinical content we actually cared about. Raw judging has hidden style bias, while style-normalized judging has normalizer bias; for our use case, the second one has been much easier to reason about.

References

[1] Zheng, L. et al. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.” NeurIPS 2023.

[2] Wang, P. et al. “Large Language Models are not Fair Evaluators.” ACL 2024.

[3] Ye, J. et al. “Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge.” ICLR 2025.

[4] Liu, Y. et al. “G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment.” EMNLP 2023.

[5] Panickssery, A. et al. “LLM Evaluators Recognize and Favor Their Own Generations.” NeurIPS 2024.

[6] Li, Q. et al. “Assessing Judging Bias in Large Reasoning Models.” COLM 2025.

[7] Li, Q. et al. “Evaluating Scoring Bias in LLM-as-a-Judge.” DASFAA 2026.

[8] Ahuja, K. et al. “MEGA: Multilingual Evaluation of Generative AI.” EMNLP 2023.

[9] Mahran, M. and Simbeck, K. “Investigating Bias: A Multilingual Pipeline for Generating, Solving, and Evaluating Math Problems with LLMs.” ECAI 2025.

[10] Zhang, T. et al. “RocketEval: Efficient Automated LLM Evaluation via Grading Checklist.” ICLR 2025.

[11] Liu, Y. et al. “Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators.” COLM 2024.

[12] Zhou, H. et al. “Mitigating the Bias of Large Language Model Evaluation.” CCL 2024.

[13] Dekoninck, J. et al. “Improving LLM-as-a-Judge Inference with the Judgment Distribution.” EMNLP 2025.

[14] Min, S. et al. “FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation.” EMNLP 2023.

[15] Huang, H. et al. “An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Models are not a General Substitute for GPT-4.” ACL Findings 2025.