An Information-Theoretic Study of RLHF-Induced Uniformity in Large Language Model Outputs

Nolan Chai* **Tianqi Zhang*** **Alex Warstadt**
University of California San Diego
{nochai, tiz019, awarstadt}@ucsd.edu

Abstract

Reinforcement Learning with Human Feedback (RLHF) is a common post-training procedure to align the outputs of Large Language Models (LLMs) with human preferences. As a result, one might expect RLHF to induce some elements of human-like audience design into LLMs. However, RLHF and other post-training alignment methods have many complex effects on the outputs of LLMs that have yet to be studied quantitatively. We apply an information-theoretic lens to investigate the changes in the "naturalness" of language and the presence of audience design in LLMs before and after post-training. The *Uniform Information Density (UID) Hypothesis* posits that humans optimize language production and comprehension across a noisy channel by transferring information at a more uniform rate. Accordingly, we analyze and compare how information is distributed within model- and human-generated text from different domains. We find that pretrained and post-trained LLMs both show superhuman uniformity across various text domains, and both RLHF and other post-training methods reduce uniformity slightly from their pretrained counterparts. However, RLHF uniquely encourages lower variance in uniformity between documents, potentially demonstrating that training on human preferences encourages consistency in information flow.

1 Introduction

A large amount of online text consumed in our daily lives has been either entirely generated by or written with the assistance of LLMs. Since early 2023, there has been a marked increase in LLM-generated text in active web pages (Spennemann, 2025), scientific writing (Liang et al., 2024), and Wikipedia articles (Brooks et al., 2024). LLMs have even been argued to pass the famous Turing Test, with humans often failing to distinguish short LLM-generated dialogues from human ones (Jones and Bergen, 2025). Yet LLM outputs still differ in many ways from human speech and writing (Guo et al., 2024; Giulianelli et al., 2023; Muñoz-Ortiz et al., 2024), suggesting a divergence in human and LLM production strategies. This raises the question: Do LLMs optimize over the same factors considered by human speakers when producing language?

In human language processing, the Uniform Information Density (UID) hypothesis holds that humans strive to maintain an even distribution of information across an utterance during language production, in order to facilitate speaker production and listener comprehension across a noisy channel (Jaeger and Levy, 2006). While LLMs lack explicit audience design mechanisms, they learn about the distributional properties of human text through pretraining on next-token prediction. They obtain a further learning signal about abstract human preferences through alignment techniques such as RLHF (Kaufmann et al., 2023). Whether these methods lead models to implicitly regulate information rate in ways that are more or less human-like, or more or less uniform, is still unknown.

Additionally, while pretrained LMs are optimized to minimize the surprisal of the training corpus, the information properties of their generated texts can still differ significantly. Previous studies have demonstrated that texts generated by pretrained LLMs are significantly *more* uniform than comparable human corpora—reliably enough to distinguish between machine-generated and human texts on the basis of information density alone (Venkatraman et al., 2024). These results raise questions about the nature of human preferences regarding information rate during language comprehension, and about how downstream alignment techniques may affect this behavior in LLMs. Reinforcement learning from human feedback (RLHF) trains LLMs not to maximize similarity to the training distribution, but to maximize a reinforcement

learning reward based on human preference data (Kaufmann et al., 2023). Thus, how might shifting the goal of an LLM from reproduction of a text to optimizing for human preferences change how it approaches distributing information?

1.1 Hypotheses

Given that pretrained LLMs are already superhuman in their uniformity, four reasonable (not necessarily mutually exclusive) hypotheses emerge:

1. Human-like uniformity is preferable to humans, thus RLHF would reduce the uniformity of a pretrained model.
2. Lower uniformity is preferable to humans as overly uniform texts are uninteresting (Tsipidi et al., 2024), thus RLHF would reduce uniformity.
3. Higher uniformity is preferable to humans as more uniform texts are easier to comprehend, thus RLHF would further increase uniformity.
4. Human preferences don't take uniformity into account, or uniformity is outranked by other preferences/considerations, thus RLHF would not significantly change uniformity.

Hypotheses 1 and 2 are particularly interesting, as they are counterintuitive to a naïve interpretation of the UID hypothesis. Due to the superhuman uniformity observed in pretrained LLMs, RLHF would seem to have to *reduce* uniformity rather than increase it to satisfy human preferences.

In this study, we ask whether and how alignment techniques, which are optimized for listener preference, alter the information rate of LLM generations, relative to both pretrained models and human texts. Our results are consistent with hypotheses 1 and 2, but ultimately support hypothesis 4 most strongly. We find that RLHF slightly decreases information uniformity, but supervised fine-tuning on specific text domains (including instruction tuning) have a similar effect. On the other hand, RLHF alone reduces the *variance* of uniformity across different model generations, leading to greater consistency of information flow across texts.

We make the following contributions:

1. A corpus of roughly 12,000 generated texts annotated with token-level surprisal values.[1]
2. A thorough analysis of different training strategies, including RLHF, instruction tuning, and domain adaptation, and their effects on the information rate of model generations.

[1] https://huggingface.co/datasets/NolanChai/rlhf-uid-generations

2 Background

The Uniform Information Density (UID) hypothesis holds that humans optimize their production of language to transfer information uniformly across a noisy channel (Fenk and Fenk, 1980; Jaeger and Levy, 2006). While some work suggests that more uniform information rate is not *always* better (Tsipidi et al., 2024, 2025), UID has been shown to affect choices in language production across many domains of language, including phonology (Aylett and Turk, 2004), syntax (Jaeger, 2010), and discourse (Genzel and Charniak, 2002). Cross-linguistic studies have also suggested grammatical rules are optimized for UID, reinforcing its importance as a foundational property of human language and cognition (Clark et al., 2023). UID measures have also been implemented as regularizers for training, resulting in LLMs producing text with higher entropy, greater lexical diversity, and a qualitative increase in "naturalness", suggesting that consideration of information flow is important for human-like text production (Wei et al., 2021).

As noted in section 1, prior work revealed significant differences in UID between BASE (pretrained) LLMs and human-generated texts (Venkatraman et al., 2024). However, these works did not examine models with RLHF post-training on human preferences. This study attempts to cover this research gap whilst also investigating whether and how UID theory applies to human preferences, as modeled by RLHF models.

2.1 RLHF

Reinforcement Learning from Human Feedback (RLHF) is a strategy for model alignment in which reinforcement learning is used to tune a pretrained LLM to optimize abstract human preferences, as estimated by a reward model trained on human feedback of LLM outputs (Kaufmann et al., 2023). This method has been especially successful for improving LLM performance on in-context learning and instruction following, resulting in the development of more effective chatbots that are optimized for conversation rather than generation of language (OpenAI, 2022; OpenAI et al., 2024). While RLHF seems to improve safety and performance, this method can lead to an "alignment tax", wherein the diversity and natural variability of outputs is reduced (Askell et al., 2021; Kirk et al., 2024; Go et al., 2023; Lin et al., 2024).

Various studies attempt to measure the improve-

ments in LLM generations resulting from post-training. Ouyang et al. (2022) introduced InstructGPT, OpenAI's first model fine-tuned with RLHF, trained using direct feedback from human annotators on LLM outputs, including qualitative judgements of instructions, toxicity, and bias, and quantitative improvements on benchmark datasets measuring truthfulness and toxicity. Other past methods evaluate effects of RLHF on the reward models' performance (Kaufmann et al., 2023) or on the LLM's generalizability and output diversity (Kirk et al., 2024). However, these metrics do not directly measure the human-likeness of the LLM outputs or explicitly compare the outputs to human text.

3 Dataset Generation

To investigate information density patterns across human and LLM-generated text, we create parallel corpora of comparable texts produced by both. Rather than asking models to imitate human writing explicitly, we follow a minimal-intervention approach similar to previous "Turing Test" benchmarks (Liu et al., 2023; Uchendu et al., 2021). For each domain, we collect human-generated texts and then prompt LLMs to generate starting from the same initial context (typically the first sentence/few sentences). We perform a minimal amount of prompt engineering to get LLM generations that are comparable to the corresponding human-written text, seeking to reduce generation artifacts while allowing us to generate from near-identical starting points without explicitly biasing models towards human-like information flow patterns.[2]

3.1 Datasets and Prompting

To test the information density of model generations across multiple text domains, we source human-generated text from four different datasets. All datasets are in English, or we subsample the English texts only.[3] We discuss prompting strategies and their possible effects on analysis in more detail in section 6.

CNN/DailyMail. To explore UID in model completions in the domain of professional writing, we use the CNN/DailyMail dataset introduced by (Nallapati et al., 2016), which consists of news articles written by journalists from CNN and the DailyMail. Articles from CNN were written between April 2007 and April 2015, while those from the DailyMail were written between June 2010 and April 2015. We choose this dataset because the articles all predate the release of ChatGPT and the widespread use of LLMs in writing news articles, limiting data contamination.[4] The source and first sentence of each article was given to each model as past context, with no explicit prompt or instruction template. The model was then allowed to fill in the rest of the article.

WritingPrompts. To extend our analysis to the creative writing domain, we use the WritingPrompts dataset (Fan et al., 2018), a corpus containing pairs of prompts and stories written by Reddit users in the subreddit r/WritingPrompts. Each story is loosely inspired by its associated prompt. For our purposes, we ignore the prompts, and feed the first sentence of each story to the model in a similar fashion to the CNN/DailyMail dataset. Since a writing prompt could spawn multiple different stories, this completion prompting method encourages more similarity between the model-generated story and the human-generated story.

DailyDialog. We use the DailyDialog dataset (Li et al., 2017) to test the uniformity of model generations in dialog completions, consisting of multi-turn, human-to-human dialog designed to reflect everyday communication, and manually transcribed to limit noise. Each dialogue d consists of a sequence of turns $d = \{t_1, t_2, ..., t_n\}$ where n represents the total number of turns in dialogue d.

For each dialogue d, we use a sliding context window approach, where our minimum context length is $k_{min} = 5$ to ensure sufficient dialogue history. For $i > k_{min}$ turns, we created multiple prompts by having incremental sliding windows. For each prompt, we extracted the dialogue up to turn i, where i is an increasing odd number from k_{min} to the total number of turns in our dialogue. Our set P of prompts $P = \{p_1, ..., p_2\} \in P$ can be represented as:

$$\{\{t_1, ..., t_5\}, \{t_1, ..., t_7\}, \{t_1, ..., t_9\}, ...p_n\}$$

[2]One notable exception was the Llama 2 7b 32k Instruct model. Examples and explanations are included in Appendix A.

[3]While the datasets we use were originally intended for tasks such as text summarization, sentiment analysis, etc, we use them here as comparable, human-generated text.

[4]It is possible, and even highly likely, that this data was used in the training of the models used in these experiments. However, it is more important in our case to avoid the inclusion of LLM-generated or LLM-assisted text in our human-generated data to avoid misconceptions about natural human uniformity.

The model is given each dialogue stub, and allowed to complete the rest of the dialogue (with no explicit prompting). Finally, all the generations from all stubs of a dialogue are combined to represent the model-generated dialogue.

WildChat. Finally, to test the UID of model outputs in a human-chatbot dialog environment, we used the WildChat dataset (Zhao et al., 2024), which consists of full conversations between users and ChatGPT. While multiple languages exist in the dataset, only English language prompts were used. WildChat differs from the others in that there is no "human-generated text" to compare to. The motivation for including this dataset is to compare the UID of model responses in the above domains to the UID in response to diverse human prompts that were meant for LLMs.

3.2 Models

We prepare generations from various language models, categorized into base, instruction tuned, domain-adapted to specific domains, and chat (RLHF) models. For our first experiment, we compare base, instruction-tuned, and RLHF models from the Llama 2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023; Zheng et al., 2024) families of models.[5]

BASE models. As a baseline, we generate completions with the base versions of each model family, trained on next-token generation alone. In our first experiment, these models are used out-of-the-box, without further fine-tuning.

RLHF models. To analyze the effect of RLHF on model UID, we used models fine-tuned using RLHF (see section 2 for more on RLHF). We choose the RLHF fine-tuned versions of the same BASE models as above.

INSTRUCTION-TUNED models. INSTRUCTION-TUNED models are LLMs fine-tuned on corpora of instruction-output pairs. This is done to improve the LLM's ability to follow instructions from a user and to adapt to a variety of tasks in-context. To preserve comparability, we use instruction-tuned versions of the same BASE models used above for Experiment 2.

[5]The specific models from the Mistral family are Mistral 7b v0.1, Mistral 7b Instruct v0.1, and Mistral Plus 7b. The models from the Llama family are Llama 2, Llama 2 7b, Llama 2 7b 32k Instruct, and Llama 2 7b Chat.

3.3 UID Calculation

For each text, we first calculate token-level surprisal. Surprisal, sometimes called the Shannon information (Shannon, 1948), is defined as the negative log-probability. We measure the surprisal of each token, conditioned on some previous context window. We estimate conditional probabilities using GPT-2 (Radford et al., 2019) with a context size of 1024 tokens. Commentary on this method can be found in section 6, UID Calculation.

$$I(w_i) = -\log_2 P(w_i|w_{<i}) \tag{1}$$

We the UID of the generated texts using three classes of metrics, following Meister et al. (2021) and Venkatraman et al. (2024):

Mean Surprisal. Mean surprisal is the average information content per token in a document $\vec{w}$:

$$\mu_{surprisal}(\vec{w}) = \frac{1}{|\vec{w}|}\sum_{i=1}^{|\vec{w}|} I(w_i). \tag{2}$$

While not itself a measure of UID, it nevertheless can be analyzed to demonstrate the tendencies of a generation method in terms of information content. In this case, $|\vec{w}|$ is the size of the document, meaning the number of tokens in the document.

Local Variance. Local variance, defined by Wei et al. (2021), measures the average change in surprisal between every pair of adjacent tokens w_{i-1} and w_i in a document $\vec{w}$, measured by some distance function $\Delta(x_1, x_2)$ (see Equation 5):

$$\text{UID}^{-1}_{pair}(\vec{w}) = \frac{1}{|\vec{w}|}\sum_{i=2}^{|\vec{w}|} \Delta(I(w_{i-1}), I(w_i)). \tag{3}$$

A document is considered uniform if it has a lower average pairwise distance, meaning it has consistently small changes in surprisal going from one token to the next. This metric aligns with optimizing for locally smooth information contours.

Surprisal Variance. Surprisal variance measures the mean distance between the surprisal of each token w_i in a document $\vec{w}$ and the mean surprisal of that document $\mu_{surprisal}(\vec{w})$, according to a distance function $\Delta(x_1, x_2)$:

$$\text{UID}^{-1}_{variance}(\vec{w}) = \frac{1}{|\vec{w}|}\sum_{i=2}^{|\vec{w}|} \Delta(I(w_i), \mu). \tag{4}$$

A document is considered uniform if it has low variance in surprisal, meaning the surprisal values of all words in the document are close to the mean surprisal of the document. Surprisal variance fits optimizing for an overall information rate, rather than local variance in information.

Distance Function. We use the Squared Difference function for Δ, following (Meister et al., 2021):

$$\Delta(x_1, x_2) = (x_1 - x_2)^2. \quad (5)$$

4 Experiment 1 - Instruction-tuning and RLHF

We hypothesize that the process of RLHF confers some sense of audience design to the model through human preference feedback, thus influencing the information density of its outputs. In our first experiment, we test this by comparing the uniformity of generations across RLHF and BASE models.

4.1 Methods

We sample 300 human-generated documents from each dataset, and extract prompts using the described strategies in subsection 3.1. Each prompt is passed to each model for generation. In total, 300 documents are generated by each model per dataset, for a total of 1200 documents per model. Outliers and empty generations are removed from consideration.[6] The human sources used to generate each prompt are saved for all datasets except for WildChat, totaling 900 human-generated documents. Then, we calculate mean surprisal, surprisal variance, and local variance for each document using the equations from subsection 3.3.

4.2 Results

All model generations have lower mean surprisal than human texts overall, indicating that models typically generate more stereotypical texts than humans; full summary statistics are reported in Appendix F (Table 6).

Model Class Analysis. Summary statistics for surprisal variance and local variance are shown in Table 1, and the distributions of both metrics are shown in Figure 1. Across both Llama and Mistral families and for both metrics, we make two observations. First, *while RLHF seems to lead to less uniform information density* (as indicated by higher median and mean values of variance for RLHF models relative to base models from the same family), these differences are small and do not pass tests of significance ($p > 0.05$, see Appendix G). Second, *RLHF leads to more consistent levels of uniform information density across documents*. Within each family, the RLHF model had a lower standard deviation for both metrics than its corresponding BASE model. The exact distributional effects vary between the two model families. For the Llama family, IQR significantly decreased ($p < 0.01$) from BASE to RLHF, while for Mistral model, the effect was not significant. However, the Mistral family exhibited significant differences in overall distribution, while the Llama family did not (see Table 7 and Table 10).

Figure 2 breaks these comparisons down by dataset, aggregating across model families. Across all datasets, the human-model relationship is striking: human texts are substantially less uniform, no matter the text domain. Additionally, the effect of RLHF is more clear on some datasets than others, though the trend is generally consistent. The local variance of RLHF models in the DailyDialog and WritingPrompts datasets are clearly higher than that of the BASE models, indicating less uniform (and thus more human-like) information contours.

The relative changes in consistency of information contours also differs slightly between domains and metrics, but variance of uniformity is still generally decreased as a result of RLHF.[7]

Model Family Analysis. Differences in center between model families are minimal in both metrics, as seen in Figure 1. However, across metrics and for all model classes, models of the Llama family had a slightly higher variance in uniformity than Mistral.

4.3 Discussion

Based on the consistent decrease in uniformity seen across both families, the effects of RLHF could possibly align with hypotheses 1 and 2. RLHF tends to result in slightly less uniform information rate compared to the base models. If this is true, several explanations are possible: First, per hypothesis 1, UID in model-generated texts is already consistently higher than in human texts, so if humans prefer more human-like texts, then RLHF should

[6] More on outlier removal can be found in Appendix B.

[7] More details on slight differences can be found in Appendix C.

	Surprisal Variance			Local Variance		
Model	**Median**	**Mean**	**Std**	**Median**	**Mean**	**Std**
Human Texts	17.483	18.656	4.793	33.797	35.633	9.027
Llama Base	**11.930**	13.642	6.494	**21.711**	**25.244**	13.543
Llama RLHF	12.493	**13.462**	**5.141**	24.301	26.907	**11.487**
Mistral Base	**12.217**	**13.279**	5.440	**21.917**	**24.570**	12.551
Mistral RLHF	12.544	13.622	**5.063**	23.822	26.003	**10.006**

Table 1: Summary statistics across all documents for both surprisal variance and local variance. Higher values mean less uniformity. In almost all cases, uniformity and std. of uniformity is lower for RLHF models.

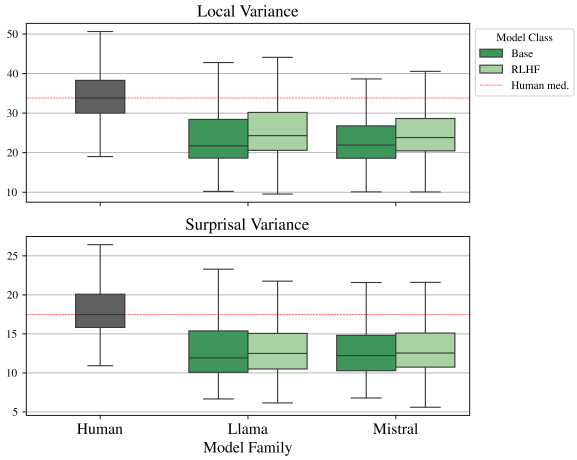

Figure 1: UID metrics for each model family, aggregated across all datasets and compared to human-UID.

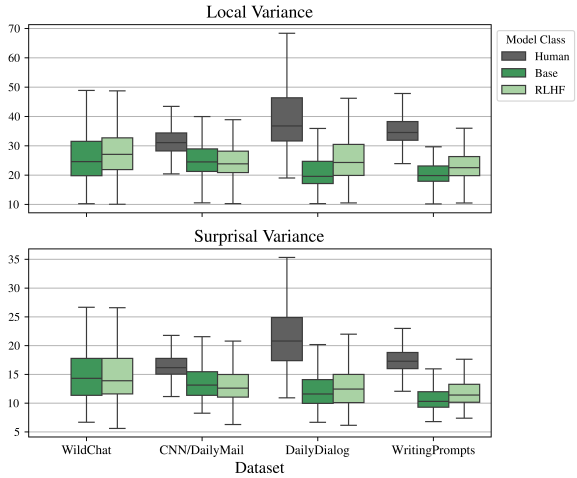

Figure 2: UID metrics for each dataset and model class, aggregated across model families.

decrease UID. Second, per hypothesis 2, UID is in tension with other desirable properties which human annotators might prioritize, such as brevity, interest, etc. These hypotheses are not mutually exclusive, as human-like levels of uniformity may be the result of jointly optimizing uniformity alongside these other considerations.

However, while the shifts in uniformity caused by RLHF are directionally consistent, they are small, failing tests of significance ($p > 0.05$). This, in turn, could possibly point towards hypothesis 4, which is that other factors outweigh uniformity and RLHF has no significant effect. As the data is extremely non-normal, though, standard tests of significance should be interpreted cautiously. Overall, the results are so far inconclusive about the effect of RLHF on the center of uniformity distributions. We further investigate these results in section 5.

However, there is a stronger and more consistent effect on the spread of uniformity distributions. The RLHF models generate texts with more consistent UID scores than the base models, with a lower standard deviation in UID metrics. This suggests that RLHF models regulate productions to stay within the same general neighborhood of information rates. Interestingly, this mirrors the low variance seen in human productions, indicating that RLHF models may implicitly control for the *consistency of information rate patterns* in a similar way to humans. Domain-specific analysis reveals that despite a general decrease in variance, slight differences exist across datasets and model families, suggesting that the effects of RLHF may vary across domains.

5 Experiment 2: Domain Adaptation and Audience Design

Experiment 1 indicated that RLHF slightly decreases information rate uniformity relative to base models, while also reducing inter-document variability. In particular, the remaining question is:

Are these effects caused by preference-based post-training, or are they due to a domain shift induced by supervised fine-tuning (SFT) methods either on chat domain data resembling the RLHF training domain or on other text domains?

To disentangle these, we compare RLHF against supervised domain adaptations of the same base model. Accordingly, we propose two competing hypotheses: **(i)** Domain shift (instruction tuning or domain adaptation either in general or on chat-domain texts) impacts information density patterns in similar ways, and **(ii)** RLHF induces distinct changes in information rate that cannot be fully explained by domain shift.

5.1 Methods

Starting from Llama-2-7B base, we train LoRA adapters for each domain (WildChat, CNN/DailyMail, DailyDialog, WritingPrompts). We merge each adapter into the base checkpoint and quantize to 8-bit GGUF for inference with "llama.cpp" (Full fine-tuning details in Appendix "Fine Tuning"). For our instruct models, we use the INSTRUCTION-TUNED checkpoints from the Llama model family (Llama 2 7b 32k Instruct).

Comparisons. We reuse the datasets and surprisal calculation methods from Experiment 1. For each evaluation dataset, we compare four model classes: (1) BASE models (no fine-tuning), (2) INSTRUCTION-TUNED models (general instruction following), (3) DOMAIN-ADAPTED models (trained on target domain), (4) CROSS-DOMAIN FINE-TUNED models (trained on other domains).

5.2 Results

Summary statistics for Llama models fine-tuned and tested on all domains, for both variance metrics are shown in Table 2. These results indicate that the DOMAIN-ADAPTED models and CROSS-DOMAIN FINE-TUNED models tended to generate slightly less uniform texts than their BASE counterparts. This matches the effect seen in RLHF models in Table 1. This effect is not unique to models finetuned on dialogue domains (WildChat and DailyDialog), but is in fact relatively consistent across most domains. However, unlike RLHF models in Experiment 1, there is no evidence for a reduction in variance due to domain adaptation, as seen in Figure 3. Humans still tended to be less uniform than DOMAIN-ADAPTED and CROSS-DOMAIN FINE-TUNED models both within individual datasets (Figure 3) and when aggregated across datasets (Figure 4), which is consistent with our results from Experiment 1.

Surp. Var.	**WC**	**CNN**	**DD**	**WP**
Human Texts	N/A	16.17	20.80	17.28
Llama Base	14.92	14.06	**10.74**	**10.38**
Llama RLHF	13.58	**11.95**	13.20	11.66
Llama Instruct	**13.20**	12.34	11.75	11.19
Llama WC	14.23	14.49	13.05	13.07
Llama CNN	15.07	12.42	14.38	13.39
Llama DD	13.84	14.99	12.84	12.17
Llama WP	14.33	13.01	13.91	12.01
Local Var.	**WC**	**CNN**	**DD**	**WP**
Human Texts	N/A	31.09	36.76	34.52
Llama Base	**24.54**	25.80	**18.56**	**20.18**
Llama RLHF	26.57	23.14	26.45	23.67
Llama Instruct	25.19	23.50	21.19	22.42
Llama WC	25.61	26.88	23.94	25.48
Llama CNN	26.57	**21.98**	25.93	24.24
Llama DD	24.83	27.26	22.42	24.38
Llama WP	24.71	24.87	25.05	22.55

Table 2: Median values for local variance and surprisal variance across fine-tuned models and datasets. Lowest values are bolded.

5.3 Discussion

We observe that DOMAIN-ADAPTED models exhibit a similar decrease in uniformity as RLHF models in Experiment 1. These results reframe our understanding of the results from Experiment 1, and could possibly support hypothesis 4, showing that for RLHF models, uniformity of information density is outranked by other factors influencing human preferences. This is because fine-tuning models on non-preference-related data reproduced even more significant shifts than Experiment 1 (see Appendix G), suggesting that the main driver was not heightened consideration for human preference, but rather a data domain shift. One possible explanation is that preferences are more sensitive to factors other than uniformity. Another could be that other drivers of human preferences, such as length or tone, are easier for the model to pick up on during training than information uniformity.

Between RLHF and domain adaptation, there *were* differences in the effect on variance in uni-

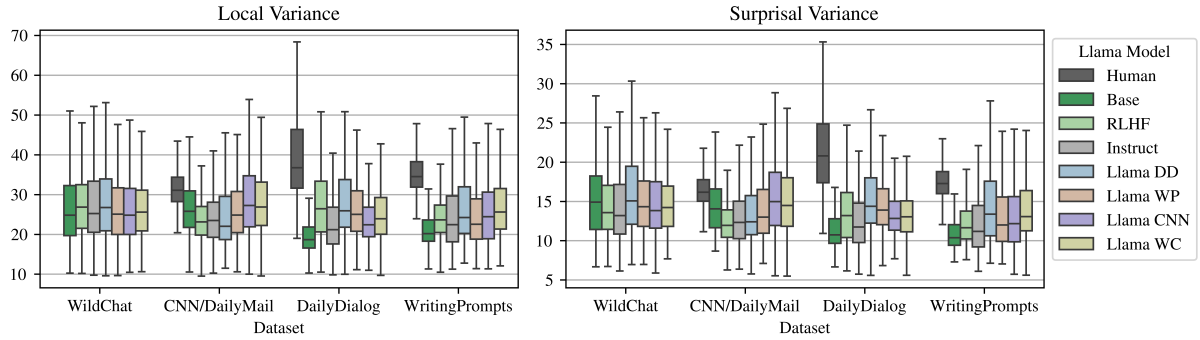

Figure 3: UID metrics for Llama models (including fine-tuned models) across datasets.

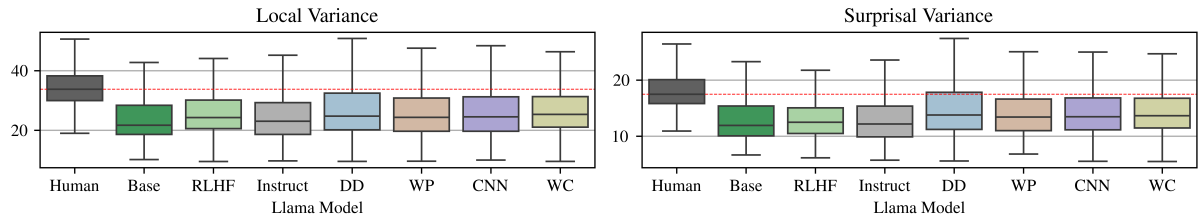

Figure 4: UID metrics for Llama models (including fine-tuned models), aggregated across datasets.

formity. This is especially evident in the surprisal variance, as shown in Figure 4 - while RLHF has constrained the variance, other fine-tuning has not. Our results show that domain adaptation does not increase consistency across documents in information flow when comparing to their base counterparts. This could suggest that human preferences may push for a more *consistent* rate of information across different generations.

6 Conclusion

Our study investigated how different fine-tuning paradigms – particularly, supervised fine-tuning and Reinforcement Learning with Human Feedback (RLHF) – affect the distribution of information in language model outputs, inspired by the Uniform Information Density (UID) hypothesis. RLHF did not increase information uniformity. Rather, it constrained the range of UID patterns a model produces by reducing variance across generated texts, while slightly lowering central tendencies relative to the same from the BASE model. However, this does not validate hypotheses 1 and 2, as other fine-tuning approaches yield similar results without optimizing human preferences. This may be due to consideration of other factors that outweigh higher uniformity in human preferences. Another possible explanation is that other factors are easier for RLHF models to learn than control of information contours, leading to little to no change. However, the lack of an increase in uniformity does suggest that, to the extent that higher uniformity is associated with greater processing ease, there is a ceiling for this effect, above which humans prioritize other factors, and pretrained LLMs have reached and possibly exceeded this ceiling. The reduction of variance also shows that RLHF models implicitly control for the variance in information patterns, similarly to humans, potentially demonstrating more precise control of information contours. Our domain adaptation experiments revealed that the effects of RLHF on the distribution of information rate are not replicable by simply fine-tuning on a dialogue domain or any other domain, thus training on human preference exhibits some special effect on the consistency of uniformity.

Overall, we corroborate earlier findings that modern LLMs exhibit higher information uniformity than human-authored text across domains, and demonstrate that training on human preferences seems to have no further effect on the uniformity of model generations. Additionally, the process of training on human preferences seems to induce a finer level of control over information flow, decreasing the variance in uniformity across different generated texts. However, these distributional shifts do not come close to human-like uniformity, suggesting that, although humans tend to author texts with similar levels of information uniformity, uniformity may not strongly impact human prefer-

ences as modeled through the process of RLHF.

Limitations

Model Prompting In this paper, we adopt specific prompting strategies to encourage the model to produce comparable generations without explicitly instructing it to generate human-like texts. We devise these prompting strategies heuristically, and we do not conduct a comprehensive comparison of strategies and their overall effect on information rate. Minor prompt changes have been shown to significantly affect LLM performance in different tasks (He et al., 2024; Ngweta et al., 2025). Therefore, future work could address this limitation by measuring the effect of giving each model more explicit instructions, rather than providing it with context and allowing it to continue the remaining documents, as was done for the CNN/DailyMail and WritingPrompts datasets.

Language limitations As mentioned in subsection 3.1, we use only English-language datasets in our analysis. While studies have upheld the UID hypothesis cross-linguistically (Clark et al., 2023), the behavior of LLMs in different languages could differ.

UID Calculation We calculate UID using GPT-2 surprisal values, following Venkatraman et al. (2024). We chose GPT-2 partly because it has higher predictive power for human reading times than larger LMs (Oh and Schuler, 2023; Lopes Rego et al., 2024), which we take to justifying its use as an estimator for human cognitive load. Additionally, prior work suggests that UID metrics computed with LM surprisals correlate more strongly with human reading times than raw frequency-based metrics (Meister et al., 2021). However, each model has its own predictions for next-token probability. It is possible that the internal measure of information rate of each model differs from the estimation according to GPT-2's probability measures. To verify for robustness, as GPT-2 may overweight frequent tokens relative to larger models, we replicated our methods using surprisals computed by a more powerful LM, Qwen (Yang et al., 2025). We found that for RLHF models, this change makes no qualitative difference in the results. The direction of change in UID metrics is still the same, although the absolute magnitude of the metrics was different. However, direction of change in variance was different for the Llama family. We therefore conclude that in the majority of cases, RLHF seems to make texts slightly less uniform, and that this result is robust to surprisal calculation. Robustness of the result of reduced variance may warrant further analysis. More detail can be found in Appendix E. However, there are certainly still limitations in using LMs for estimating information rate. Biases have been observed in LMs as models for human cognitive behavior (Haller et al., 2024). Future work could seek to establish best practices for estimating information rate.

Fine-Tuning Due to computational constraints, we use 8-bit quantizations of the models from GGUF and llama.cpp, and fine-tune them using low rank adaptation (LoRA). LoRA allows us to specialize Llama-7B cheaply, but its low-rank updates updates only a fraction of the network. We may have observed larger effects due to fine-tuning if we had used full fine-tuning or full-precision models.

Acknowledgements

We would like to thank Dr. Mario Giulianelli for his helpful feedback during the formulation of this paper.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, and 3 others. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Creston Brooks, Samuel Eggert, and Denis Peskoff. 2024. The rise of ai-generated content in wikipedia. *Preprint*, arXiv:2410.08044.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for uniform information density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Ralph D'Agostino and E. S. Pearson. 1973. Tests for departure from normality. empirical results for the

distributions of b2 and b1. *Biometrika*, 60(3):613–622.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

August Fenk and Gertraud Fenk. 1980. Konstanz im Kurzzeitgedächtnis –Konstanz im sprachlichen Informationsfluß? *Zeitschrift für experimentelle und angewandte Psychologie*, 27(3):400–414.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. 2023. Aligning language models with preferences through f-divergence minimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking linguistic diversity of large language models. *Preprint*, arXiv:2412.10271.

Patrick Haller, Lena S Bolliger, and Lena A Jäger. 2024. On language models' cognitive biases in reading time prediction. In *ICML 2024 Workshop on LLMs and Cognition*. University of Zurich.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *Preprint*, arXiv:2411.10541.

Tim Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Tim Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Cameron R. Jones and Benjamin K. Bergen. 2025. Large language models pass the turing test. *Preprint*, arXiv:2503.23674.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 10.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of rlhf on llm generalisation and diversity. *Preprint*, arXiv:2310.06452.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the increasing use of llms in scientific papers. *Preprint*, arXiv:2404.01268.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606, Miami, Florida, USA. Association for Computational Linguistics.

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2023. Check me if you can: Detecting chatgpt-generated academic writing using checkgpt. *arXiv preprint arXiv:2306.05524*, 2.

Adrielli Tina Lopes Rego, Joshua Snell, and Martijn Meeter. 2024. Language models outperform cloze predictability in a cognitive model of reading. *PLOS Computational Biology*, 20(9):1–24.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10).

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. 2025. Towards LLMs robustness to changes in prompt format styles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 529–537, Albuquerque, USA. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

OpenAI. 2022. Introducing ChatGPT.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff 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. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

C. E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Dirk HR Spennemann. 2025. Delving into: the quantification of ai-generated content on the internet (synthetic data). *Preprint*, arXiv:2504.08755.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Eleftheria Tsipidi, Samuel Kiegeland, Franz Nowak, Tianyang Xu, Ethan Wilcox, Alex Warstadt, Ryan Cotterell, and Mario Giulianelli. 2025. The harmonic structure of information contours. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31636–31659, Vienna, Austria. Association for Computational Linguistics.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform Information Density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18820–18836, Miami, Florida, USA. Association for Computational Linguistics.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. GPT-who: An information density-based machine-generated text detector. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Clara Meister, and Ryan Cotterell. 2021. A cognitive regularizer for language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5191–5202, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *Preprint*, arXiv:2405.01470.

Chen Zheng, Ke Sun, Hang Wu, Chenguang Xi, and Xun Zhou. 2024. Balancing enhancement, harmlessness, and general capabilities: Enhancing conversational llms with direct rlhf. *Preprint*, arXiv:2403.02513.

Appendix

A Llama Instruct Prompting

With the standard prompting strategies, the Llama 2 7b 32k Instruct model produced artifacts such as special instruct tokens, chat template tokens, etc. in its generations. In order to clean up generations, the Llama 2 chat template was applied to every prompt with an instruction preceding the text snippet. Special strings such as `[INST]` and `[/INST]` were added as stop strings, such that the model generation was halted upon observation of these strings. Unless otherwise specified, all models including chat and instruct variants were queried in pure continuation mode; we supplied only the partial document context and terminated generation as mentioned above, with no special system or user instruction for terminating generation.

B Outlier Removal

In our dataset generation, we tried to remove as many unreasonable generations as possible through a minimal prompt engineering process. However, there still remained texts that had unreasonable surprisal values, leading to extremely low or high uniformity. Qualitative assessment of these outliers revealed that many were nonsensical generations or, in the case of many WildChat generations, not in English. This led to the generation of tokens that had extreme surprisal values, such as characters in another language or programming language syntax. Some prompts also led to empty generations, from which a uniformity calculation would be impossible.

To clean our dataset, we removed any documents that displayed above two times the human maximum or below one-half the human minimum for either of the two uniformity metrics, including empty generations. This was done with consideration of our overall motivation of concerns over unnaturalness in LLM-generated content. If a human were trying to generate, say, a news article with an LLM, such outliers would immediately stand out to them and be discarded. Removing such outliers reduced our total number of generations from 12,000 to 11,674. On average, less than 3% of texts were removed, and the distribution of removed texts was even across models and datasets. Much of the analysis was unchanged, but this corrected for over-estimations of variance in uniformity for the INSTRUCTION-TUNED models in particular.

Surp. Var.	**WC**	**CNN**	**DD**	**WP**
Human Texts	N/A	2.06	6.46	2.49
Llama Base	8.92	3.94	7.11	**3.34**
Llama RLHF	**7.74**	**3.19**	**4.25**	3.68
Mistral Base	8.40	**3.12**	4.31	3.64
Mistral RLHF	**7.45**	3.71	**4.16**	**2.41**
Local Var.	**WC**	**CNN**	**DD**	**WP**
Human Texts	N/A	4.43	11.66	6.10
Llama Base	18.47	7.60	16.13	**6.87**
Llama RLHF	**16.94**	**6.68**	**10.83**	7.68
Mistral Base	20.66	**6.40**	9.74	7.66
Mistral RLHF	**15.01**	7.25	**8.71**	**4.62**

Table 3: Standard Deviation of surprisal variance and local variance across models and domains.

C Variance in Uniformity

Standard deviations across models and domains for surprisal variance and local variance are shown in Table 3. In most cases, RLHF reduces the variance in uniformity, with the exceptions of Writing-Prompts for the Llama family and CNN/DailyMail for the Mistral family.

D Fine Tuning

For each target domain (news, human-human & human-chatbot dialogue, creative writing), we collected $n > 2000$ documents, cleaned whitespace and removed instances shorter than 50 characters. Datasets were shuffled and split 80/10/10 into train, validation, and test sets.

We fine-tuned the **Llama-2-7B** base checkpoint, with the following configuration for parameter-efficient updates via LoRA (low rank adaptation):

- Target modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj.
- Rank $r = 24$, $\alpha = 48$, dropout = 0.05.

The hyperparameter configurations for each fine tune are listed in Table 4.

For tokenization, we used the HuggingFace Llama-2 tokenizer and default settings. We performed a heuristic search before a grid search over smaller parameter ranges to optimize for hyperparameters on perplexity. For inference, we merged the LoRA adapters onto the Base GGUF weights before converting to an 8-bit quantization,

Batch Size	8
Grad Accumulation	4
Epochs	5
Optimizer	AdamW (fused)
Learning Rate	2×10^{-5} (cosine decay)
Warm-up	10% of steps
Weight Decay	0.01
FP16	Enabled
Early Stopping	Patience=3 eval steps

Table 4: Hyperparameters used for each fine tune.

using the same generation parameters as the base model. Here were our perplexity scores for base versus our domain fine-tunes:

Dataset + Perplexity	Base	Fine-tune
Daily Dialog	7.684	3.698
WildChat	4.109	2.931
CNNDailyMail	13.911	5.597
WritingPrompts	11.288	9.829

E Qwen Experiments

In order to verify the robustness of our results to a different surprisal calculation method, we replicated our methods using surprisals computed by a more powerful LLM, Qwen2.5 (Yang et al., 2025). Using the WritingPrompts dataset, we performed experiment 1 (as described in subsection 4.1) and experiment 2 (as described in subsection 5.1, focusing on the Instruct fine-tuned models). We calculated the same metrics with Qwen2.5, comparing them to our original's GPT-2 in subsection 3.3. We then aggregated the data into median and standard deviation, as shown in Table 5.

While changing the model affected the overall magnitude of the changes, the results remain the same. RLHF models are slightly less uniform and have less variance than BASE models. SFT models exhibit similar shifts in median uniformity, but have much higher variance.

F Mean Surprisal

For Experiment 1, we also calculated the mean surprisals of texts generated by each model. This in itself does not give any information about uniformity and is heavily affected by the length of the text. However, similar relationships can be seen. Human texts are more surprising on average than LLM-generated texts, and the RLHF models have lower variance in mean surprisal for both model families.

	Surprisal Variance		Local Variance	
Model	**Median**	**Std**	**Median**	**Std**
Llama Base	**12.64**	**4.73**	**20.18**	**6.87**
Llama Instruct	14.58	12.80	22.46	12.59
Llama RLHF	15.21	6.57	23.67	7.68
Mistral Base	12.06	4.69	**19.54**	7.66
Mistral Instruct	**11.75**	7.82	20.07	10.75
Mistral RLHF	14.62	**4.27**	21.76	**4.62**

Table 5: Results from Qwen2.5 on WritingPrompts

Model	Median	Mean	Std
Human Texts	**4.83**	**4.93**	0.80
Llama Base	4.01	4.02	0.94
Llama Instruct	3.77	3.85	**1.08**
Llama RLHF	3.79	3.93	0.88
Mistral Base	4.00	4.07	0.87
Mistral Instruct	3.72	3.95	1.07
Mistral RLHF	4.05	4.15	0.74

Table 6: Summary statistics of surprisals of documents across models. Human texts had the highest mean, while Llama 2 7b 32k Instruct had the highest variance.

G Statistical Tests

Using the difference of means of surprisal variance as our test statistic, we ran paired t-tests ($H_0 : \mu_{\text{Model A}} = \mu_{\text{Model B}}$, $H_a : \mu_{\text{Model A}} \neq \mu_{\text{Model B}}$, $\alpha = 0.05$) to evaluate some of our previous conclusions. As seen in Table 9, the test revealed that we fail to reject the null for differences between base and RLHF models, supporting the conclusion that RLHF has little to no effect on information rate. In comparing supervised fine-tuned (SFT) models to base models, the Llama domain-adapted models were found to be significantly different from the base. However, the Llama instruct model was not.

To test for normality, we ran D'Agostino-Pearson tests (D'Agostino and Pearson, 1973), which reject normality for all models (all $p \leq 10^{-6}$). Results are shown in Table 8. While humans displayed a skew of 2.66 and excess kurtosis of 12.75, models exhibit very heavy tails (skew often ≤ 10, excess kurtosis 100 to 560). This suggests that t-test results may be slightly off, as our distributions were not entirely normal.

To test for the effect of RLHF on shape and spread of the distributions, we ran Mann-Whitney U tests (Mann and Whitney, 1947). Results can be seen in Table 10. These found significant differences across all distributions except for between Llama base vs Instruct, indicating that while center had not changed, the distributions were still different after most fine-tuning strategies had been applied. Further analyzing the effect on spread, we computed IQRs, comparing RLHF to base counterparts, and SFT to base. Finding that the RLHF models had lower IQRs than base models, we tested the significance of the IQR difference with permutation tests ($H_0 : IQR_{base} = IQR_{RLHF}$, $H_a : IQR_{base} \neq IQR_{RLHF}$, $\alpha = 0.05$). As shown in Table 7, we found that while the Llama RLHF model had a significant reduction in IQR ($p < 0.01$), the Mistral model failed to reject the null ($p > 0.05$). Based on the results of the D'Agostino-Pearson tests and Mann-Whitney U tests as well as qualitative examination of the distributions, we believe RLHF caused a larger change in the tails of the distribution than the center for Mistral models. A similar effect was observed in our SFT models, which also failed the IQR test of significance ($p > 0.05$).

Overall, we observed a tightening of the distribution implied by RLHF, with tails decreasing, and a slight broadening of distributions through SFT.

Model	**ΔIQR**	p
Mistral RLHF	-0.171	0.523
Llama RLHF**	-0.743	0.007
Llama Instruct	0.185	0.519
Llama WC	-0.020	0.957
Llama CNN**	1.304	0.001
Llama DD	0.368	0.305
Llama WP	0.318	0.363

Table 7: IQR Differences (fine-tuned - base) and permutation test results. Starred comparisons exhibited a significant difference in IQR ($* : p < 0.05$, $** : p < 0.01$).

Model	**Skew**	**Kurt.**	p
Human**	2.66	12.73	<0.01
Mistral Base**	4.71	38.03	<0.01
Mistral RLHF**	3.67	27.66	<0.01
Llama Base**	6.35	70.79	<0.01
Llama RLHF**	5.53	63.96	<0.01
Llama Instruct**	2.44	14.42	<0.01
Llama WC**	2.52	12.68	<0.01
Llama CNN**	1.96	6.47	<0.01
Llama DD**	1.61	5.44	<0.01
Llama WP**	3.52	25.78	<0.01

Table 8: D'Agostino-Pearson tests for normality of distributions. For all starred distributions, the null hypothesis is rejected and the distribution is non-normal ($* : p < 0.05$, $** : p < 0.01$).

Model	**Mean Δ**	p
Mistral RLHF	+0.31	0.09
Llama RLHF	-0.17	0.46
Llama Instruct	-0.28	0.23
Llama WC**	+1.26	<0.01
Llama CNN**	+1.78	<0.01
Llama DD**	+0.94	<0.01
Llama WP**	+0.86	<0.01

Table 9: Results of t-tests comparing surprisal variance distributions with corresponding base models, with mean pairwise differences (Base - FT). Starred models had a significant difference in means ($* : p < 0.05$, $** : p < 0.01$).

	vs. Human		vs. Base	
Model	U	p	U	p
Mistral Base*	8.92×10^5	$< .01$	N/A	N/A
Mistral RLHF**	8.95×10^5	$< .01$	7.26×10^5	.004
Llama Base*	8.70×10^5	$< .01$	N/A	N/A
Llama RLHF*	9.08×10^5	$< .01$	7.19×10^5	.066
Llama Instruct*	8.69×10^5	$< .001$	6.68×10^5	.524
Llama WC**	8.13×10^5	$< .01$	8.33×10^5	$< .01$
Llama CNN**	7.57×10^5	$< .01$	8.19×10^5	$< .01$
Llama DD**	8.03×10^5	$< .01$	7.84×10^5	$< .01$
Llama WP**	8.11×10^5	$< .01$	7.79×10^5	$< .01$

Table 10: Mann-Whitney U test results comparing distributions. Starred comparisons exhibited a significant difference in overall distribution with one (*) or both (**) corresponding base models and human data.