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

Anonymous ACL submission

Abstract

Reinforcement Learning with Human Feedback is an increasingly popular post-training procedure for Large Language Models (LLMs) to better align outputs with human values and increase output quality. As LLMs continue to be incorporated and improved for various modes of natural language communication, one might expect some sense of human-like audience design to be induced into LLMs. However, the effects of RLHF on the considerations that shape LLM text production is difficult to quantify. Thus, we propose employing an information-theoretic lens to investigate the changes in the "naturalness" of language and presence of audience design in LLMs trained using fine-tuning and RLHF methods. On the basis of the *Uniform Information Density (UID) Hypothesis*, which posits that humans optimize their production of language to transfer information uniformly across a noisy channel, we analyze and compare how information is distributed within model-generated and human-generated text belonging to various domains to investigate the presence and form of audience design in LLMs. With two primary metrics of information uniformity, surprisal variance and local consistency, we find that RLHF seems to encourage less variance in information rates across generations, while fine-tuning decreases uniformity, shifting distributions slightly in the direction of human-generated text. However, models still exhibit significantly superhuman uniformity across various domains of text. Our results reveal that while modern LLM training and fine-tuning paradigms have made progress in approximating human-like information distributions, systematic differences persist.

1 Introduction

An increasing amount of the online text we consume in our daily lives has been either entirely generated by LLMs or written with LLM-assisted tools. 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). Additionally, while humans often fail to distinguish short LLM-generated dialogues from human ones (Jones and Bergen, 2025), they are still linguistically different in many ways (Guo et al., 2024; Giulianelli et al., 2023; Muñoz-Ortiz et al., 2024). The increasing prevalence of LLM-generated texts and subtle divergence of such texts from natural human usage makes the characterization of the differences between machine outputs and human authored text increasingly crucial.

One way to approach this is by analyzing the considerations that shape text production. For human speakers, a prominent hypothesis is the Uniform Information Density (UID) hypothesis, which postulates that human producers of language strive to maintain an even distribution of information across an utterance in order to facilitate listener comprehension (Jaeger and Levy, 2006). While LLMs lack explicit audience design mechanisms, they learn from human text, and—through more modern methods such as RLHF—from abstract human preferences (Kaufmann et al., 2023). These methods could lead models to implicitly regulate information rate in ways that approximate UID, or to deviate from it in systematic ways. Additionally, while autoregressive LMs are designed to minimize mean surprisal, their spatial distribution can still differ significantly.

In this study, we adopt an information-theoretic perspective to quantify these differences between modern LLM outputs and human productions in their considerations for audience design, and how they have evolved across modern LLM training methods. Crucially, we do **not** presuppose that humans and LLMs generate text in the same way. Instead, we treat UID as a shared observable space in which different production mechanisms can be compared. In particular, we analyze the effect of

alignment techniques such as RLHF on the human likeness of their productions on the basis of the Uniform Information Density (UID) Hypothesis.

We therefore ask whether and how alignment techniques, which are optimized for listener preference, encourage LLM generations to exhibit UID-like behavior relative to human texts. We find that RLHF techniques actually have little effect on the level of uniformity, but rather reduces the *variance* of uniformity in model generations, showing greater consistency of information flow across texts. Domain adaptation through supervised fine-tuning, on the other hand, has mixed effects on variance between different text domains, but generally decreases uniformity, aligning generations closer to human 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.

2 Background

2.1 Uniform Information Density

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). 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 (Clark et al., 2023) have also suggested grammatical rules are optimized for UID, reinforcing its importance as a foundational property of human language and cognition.

Prior work examining the UID of LLM outputs reveals significant differences between base (non-RLHF) LLMs and human-generated texts. Venkatraman et al. (2024) demonstrated that text generated by LLMs is significantly more uniform than comparable human corpora- to the point that UID-based features can reliably distinguish between machine-generated and human texts. This supports that LLM pretraining induces an implicit control over information rate, yet also demonstrates that this sense results in superhuman uniformity. Models built with RLHF that train more directly on human preferences were also not studied.

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).

2.2 RLHF

Reinforcement Learning from Human Feedback (RLHF) is a strategy for reinforcement learning that incorporates abstract human preferences through 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 straight generation of language (OpenAI, 2022; OpenAI et al., 2024). Though 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). However, this is hard to measure objectively.

Various studies have made efforts to measure the improvements in the generations of the language model. 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 generalisability 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. Instead, these comparisons remain implicit, assuming that human annotators prefer more "human-like" productions.

Under the UID hypothesis, humans may engage in audience design by optimizing for a more uniform information rate in consideration of processing constraints on the comprehender (Jaeger, 2010). Thus, we hypothesize that the addition of RLHF

[1] https://github.com/anon/dataset_repo

fine-tuning to a BASE model increases the uniformity of model outputs. With RLHF fine-tuning, a model would learn similar facets of audience design from training on abstract human preferences, and thus an increase in uniformity in consideration of the end user. Due to human production constraints, an LLM would also be better positioned than a human producer to optimize its information rate for a comprehender. In this way, RLHF would diverge the model from human-like information rates, making outputs less similar to natural language from a UID perspective.

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 rigorously 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 7.

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] For prompting, 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:

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

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

[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.

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.

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.

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

3.3 UID Calculation

For each text, we first calculate token-level surprisal. Surprisal, sometimes called the Shannon information following (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 7, UID Calculation.

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

With the surprisal values, we then evaluated 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 measures the average information content in a document $\vec{w}$:

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

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, whereas $I(w_i)$ is the surprisal of the i^{th} token in the document.

Pairwise Surprisal Distance/Local Consistency. Local consistency, defined by Wei et al. (2021), measures the average change in surprisal between every pair of tokens w_{i-1} and w_i in a document $\vec{w}$, measured by some distance function $\Delta(x_1, x_2)$ (see Equation 4):

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

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 $\Delta(x_1, x_2)$:

$$\text{UID}_{variance}(\vec{w}) = \frac{1}{|\vec{w}|}\sum_{i=2}^{|\vec{w}|} \Delta(I(w_i), \mu). \quad (3)$$

[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.

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 consistency in information.

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

$$\Delta(x_1, x_2) = (x_1 - x_2)^2. \quad (4)$$

4 Experiment 1 - Instruction-tuning and RLHF

We hypothesize that RLHF confers some influence of audience design to the model through human feedback, which would increase the uniformity of its generations. We test this hypothesis by comparing RLHF and BASE models. Additionally, we ask whether there are similar effects due to the alignment of the model to more chatbot-like through instruction-tuning, or whether human feedback is unique. In our first experiment, we test this by comparing the uniformity of generations across RLHF, INSTRUCTION-TUNED, 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 consistency for each document using the equations from subsection 3.3.

4.2 Results

4.2.1 Mean Surprisal

Mean surprisal values are shown in Table 1. Mean surprisal varied greatly between models and humans. All models produced generations with lower average mean surprisal than human generations, indicating that models typically generate more stereotypical texts than humans.

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

Model	Median	Mean	Std
Human Texts	17.48	18.66	4.79
Llama Base	11.87	13.44	5.88
Llama Instruct	12.14	13.22	5.19
Llama RLHF	12.48	13.41	5.17
Mistral Base	12.17	13.21	5.46
Mistral Instruct	**11.60**	**12.75**	4.98
Mistral RLHF	12.53	13.61	**5.07**

Table 2: Summary statistics for surprisal variance. Higher values mean less uniformity.

4.2.2 UID Metrics

Model Class Analysis. Across both model families, RLHF *reduces the inter-document standard deviation* of UID scores by roughly 5–20 %, while leaving median uniformity unchanged or slightly lower, as shown by Figure 1. Table 2 and Table 3 display the standard deviations of each metric for each model. Within each family, the RLHF model had a much lower standard deviation for both metrics than its corresponding BASE model. The effect of instruction tuning was less consistent.

Figure 2 breaks these comparisons down by dataset, aggregating across model families. Across all datasets, the human-model relationship seems to hold true: human texts are, in general, less uniform, no matter the text domain. Additionally, BASE models in the dialog dataset are more more uniform than their instruction-tuned counterparts when looking at local consistency, but close to equally uniform in surprisal variance. The same is true of WildChat.

Model Family Analysis. Differences in center between model families are minimal in both metrics, as seen in Figure 1. However, across metrics

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

Model	Median	Mean	Std
Human Texts	33.80	35.63	9.03
Llama Base	**21.64**	24.82	12.89
Llama Instruct	22.89	25.12	11.06
Llama RLHF	24.22	26.78	11.56
Mistral Base	21.86	**24.42**	12.60
Mistral Instruct	22.59	25.71	13.34
Mistral RLHF	23.82	25.98	**10.01**

Table 3: Summary statistics for local consistency. Higher values mean less uniformity.

and for all model classes, models of the Llama family had a slightly higher variance in uniformity than Mistral.

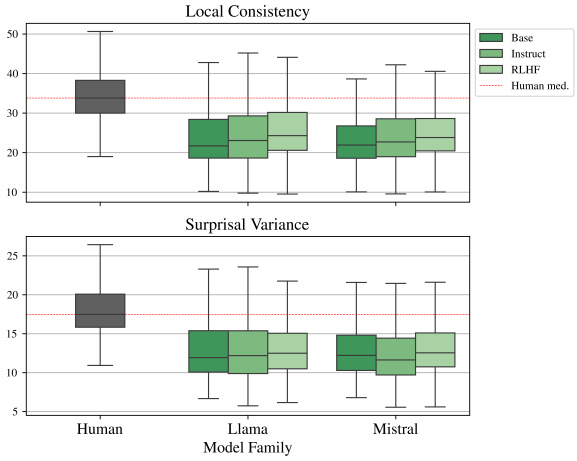

Figure 1: UID metrics for each model family, aggregated across all datasets and compared to human-UID.

4.3 Discussion

The effects of RLHF on information uniformity do not align with our initial hypothesis that RLHF would increase uniformity. If anything, RLHF tends to result in slightly less uniform information rate compared to the base models. Several explanations are possible: First, UID is in tension with other desirable properties, such as brevity, which human annotators might prioritize. Second, UID in model-generated texts is already consistently higher than in human texts, so if humans prefer more human-like texts, then RLHF should decrease UID. Or, perhaps there is a ceiling above which humans do not prefer higher UID.

When we consider all the texts generated by a model, the RLHF models generate texts with more similar 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. The effect of instruction tuning on information rate, on the other hand, is more sensitive to model family, likely due to specific datasets or strategies in supervised fine-tuning.

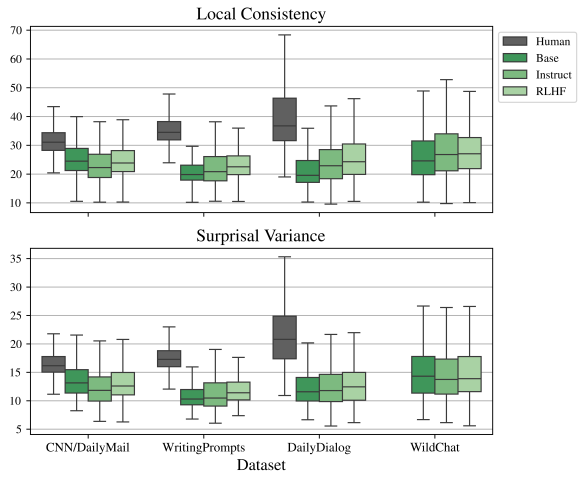

Figure 2: UID metrics for each dataset and model class, aggregated for each model family.

The results for specific domains reveal that INSTRUCTION-TUNED and RLHF models have more varied information contours than BASE models in conversational contexts (via the DailyDialog dataset), potentially reflecting more naturalistic turn-taking patterns. While these models may not fully replicate human information density patterns, these training methods produce domain-appropriate information structuring.

5 Experiment 2: Domain Adaptation and Audience Design

Our initial analysis reveals that all language models, regardless of training method, produce text with significantly higher information uniformity than human-written text. However, subtle differences between model classes warrant a deeper investigation into the role of text generation across different domains. A potential confound in the above experiments is that the effects we observe could be due to fine-tuning on the chat domain, rather than learning to optimize human preferences or perform in-context learning. To determine the impact of this potential confound, we examine variants of the

BASE models that have undergone domain-specific fine-tuning, or domain adaptation.

While our first experiment focused primarily on comparing BASE models with models, we extended our analysis to include fine-tuned models as a distinct category, exploring two competing hypotheses: **(i)** Any form of domain-specific fine-tuning (instruction tuning or domain adaptation) impacts information density patterns in similar ways, and **(ii)** RLHF and to a lesser extent instruction tuning induce distinct changes in information rate that cannot be fully replicated through other fine-tuning approaches.

5.1 Methods

For Experiment 2, we create several custom fine-tuned models based on the Llama 2 7B architectures. We fine-tune the Llama 2 7B full model on each domain before encoding the models into an 8-bit quantization, in the GGUF weights format for inference in "llama.cpp". Then, we test the four resulting fine-tuned models.

Experimental Setup. We use the same datasets and surprisal calculation methodology as in Experiment 1. For each domain (news, dialogue, creative writing), we compare the following models: (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

5.2.1 Mean Surprisal

Model	WC	CNN	DD	WP
Human Texts	N/A	4.29	**5.22**	**5.04**
Llama Base	4.09	4.11	3.82	4.03
Llama Instruct	3.50	3.91	3.72	3.94
Llama RLHF	3.49	3.51	4.27	3.93
Llama WC	3.91	4.16	4.17	4.51
Llama CNN	4.20	3.91	4.52	4.55
Llama DD	**4.10**	**4.50**	4.40	4.51
Llama WP	4.18	4.28	4.33	4.50

Table 4: Median values for mean surprisal across fine-tuned models on the WildChat (WC), CNN/DailyMail (CNN), Daily Dialog (DD), and WritingPrompts (WP) datasets.

Table 4 shows median values for mean surprisal for the models in Experiment 2 on each of the data domains. Across the domains, we fail to find consistent evidence that supervised domain adaptation meaningfully alters mean surprisal. The broader UID picture echoes this null result: fine-tuned models do not appear to have an effect in correcting for this information-rate disparity; they remain substantially more uniform than human texts, and the degree of deviation is unaffected by whether the domain is domain-matched or cross-domain.

Model	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 Instruct	**13.20**	12.34	11.75	11.19
Llama RLHF	13.58	**11.95**	13.20	11.66
Llama WC	14.23	13.09	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	13.94	12.01

Table 5: Median values for surprisal variance across fine-tuned models and datasets.

5.2.2 UID Metrics

Model	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 Instruct	25.19	23.50	21.19	22.42
Llama RLHF	26.57	23.14	26.45	23.67
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 6: Median values for local consistency across fine-tuned models and datasets.

Table 6 and Table 5 show that the DOMAIN-ADAPTED models and CROSS-DOMAIN FINE-TUNED models tended to be about as uniform, if slightly less, than their BASE counterparts, suggesting similar information rates. This matches the effect seen in RLHF models in Table 3. 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

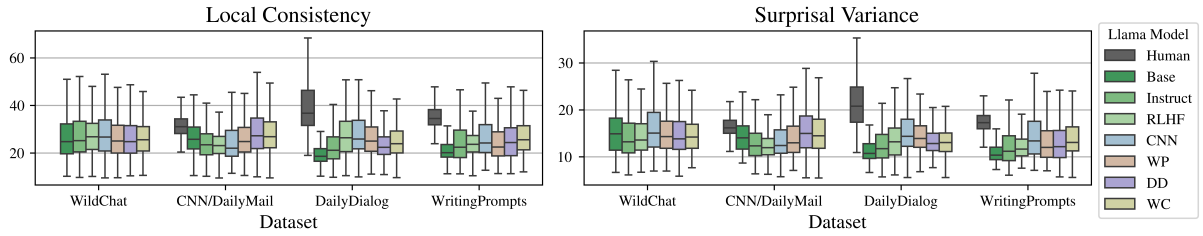

Figure 3: UID metrics for Llama models (including fine-tuned models) across datasets.

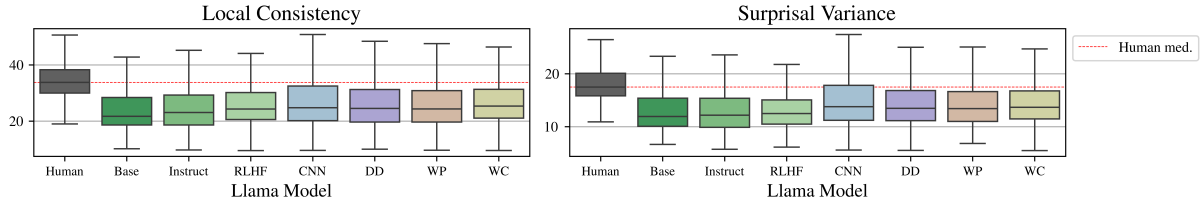

Figure 4: UID metrics for Llama models (including fine-tuned models), aggregated across datasets.

across datasets (Figure 4), which is consistent with our results from Experiment 1.

5.3 Discussion

We observe that DOMAIN-ADAPTED models exhibit a similar lack of effect on the center of uniformity distributions as RLHF models in Experiment 1. This confirms that the lack of shift in center was not due to any special effects in Experiment 1 from fine-tuning on a dialogue domain; fine-tuning in general seems to have little to no effect on the uniformity of model generations. However, there were differences in the effect on variance in uniformity. Our results show that domain adaptation does not increase consistency in information flow when comparing to their base counterparts. This suggests that human preferences may push not for more or less uniform text, but rather for a more consistent rate of information across different generations.

6 Conclusion

Our study investigated how different fine-tuning paradigms – particularly, instruction 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. Contrary to our initial hypothesis, RLHF did not uniformly increase information uniformity. Rather, it constrained the range of UID patterns a model produces by reducing variance across generated texts, while leaving central tendencies relatively the same from the BASE model, or even lowering uniformity slightly. This suggests that while RLHF tuning had little to no effect on information rate, RLHF models still implicitly control for the variance in information flow patterns, similarly to humans. Our domain adaptation experiments revealed that the effects of RLHF are not replicable by simply fine-tuning on a dialogue domain or any other domain, suggesting that training on human preference exhibits some special effect on the consistency of uniformity. Furthermore, aligning models to particular text genres typically has little effect on uniformity.

Overall, we corroborate earlier findings that modern LLMs exhibit higher information uniformity than human-authored text across domains, and demonstrate that even more modern fine-tuning paradigms have minimal impact on the uniformity of model generations. While RLHF does have an effect of information uniformity, it neither increases uniformity as predicted if the model is learning audience design principles, nor significantly decreases uniformity to human-like levels.

7 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. 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, especially for low-resource languages.

UID Calculation We calculate UID using GPT-2 surprisal values, following the practice of (Venkatraman et al., 2024). We chose GPT-2 partly because it has higher predictive power for human reading times than very large LMs (Lopes Rego et al., 2024), making it a decent estimate for human cognitive load; additionally, prior work has shown that UID metrics computed with LM surprisals predict human reading times better than raw frequency-based metrics. 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 (Qwen 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. We therefore conclude that in the majority of cases, RLHF makes texts less uniform, and that this result is robust to surprisal calculation. More detail can be found in Appendix D 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 restraints, we use 8-bit quantizations of the models through GGUF and llama.cpp, and fine-tuned using parameter-efficient methods via low rank adaptation (LoRA). LoRA allows us to specialize Llama-7B cheaply, but its low-rank updates touch on only a fraction of the network, so deeper discourse patterns and UID are not as affected as if we had used a broader architecture, longer full-precision or LoRA runs, and a loss that directly rewarded UID for fuller experimentation.

8 Potential Risks

A potential risk of this research is in guiding the obfuscation of LLM-generated text. Since we have established known disparities in uniformity between human text and model generations, work could be done to account for this in pursuit of hiding LLM use. This risk is minor at present as there are no clear techniques to produce generations with more human-like UID. It should also be noted that UID metrics are not a stand-alone authenticity metric of human natural language in the absolute. Our UID metrics were measured using GPT-2 probabilities; mismatches between that lens and real-world comprehension may mislead less informed readers.

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.

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.

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

T. 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.

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 Gu̇lçehre, 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.

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.

Qwen, :, 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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

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.

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.

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.

C 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 **??**.

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 7: Hyperparameters used for each fine tune.

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, 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

D 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 (Qwen et al., 2025). Using the WritingPrompts dataset, we performed experiment 1. Using probability values from Qwen2.5 instead of GPT2, we calculated the same metrics with the equations described in subsection 3.3. We then aggregated the data into median and standard deviation, as shown in Table 8 and Table 9.

Model	Median	Std
Llama Base	20.180481	6.866043
Llama Instruct	22.461929	12.592461
Llama RLHF	23.666833	7.681891
Mistral base	19.541096	7.664134
Mistral Instruct	20.069720	10.746450
Mistral RLHF	21.758398	4.621838

Table 8: Local Consistency calculations from GPT2 on WritingPrompts

Model	Median	Std
Llama Base	12.639551	4.727139
Llama Instruct	14.583953	12.800454
Llama RLHF	15.208221	6.567605
Mistral Base	12.058930	4.688608
Mistral Instruct	11.753679	7.816419
Mistral RLHF	14.615638	4.272996

Table 9: Local Consistency calculations from Qwen2.5 on WritingPrompts