Limits to Predicting Online Speech Using Large Language Models

Mina Remeli[1], Moritz Hardt[1], and Robert C. Williamson[2]

[1]*Max-Planck Institute for Intelligent Systems, Tübingen and Tübingen AI Center*
[2]*University of Tübingen and Tübingen AI Center*

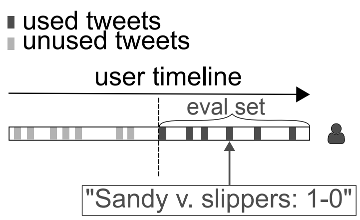
Abstract

We study the predictability of online speech on social media, and whether predictability improves with information outside a user's own posts. Recent work suggests that the predictive information contained in posts written by a user's peers can surpass that of the user's own posts. Motivated by the success of large language models, we empirically test this hypothesis. We define unpredictability as a measure of the model's uncertainty, i.e., its negative log-likelihood on future tokens given context. As the basis of our study, we collect a corpus of 6.25M posts from more than five thousand X (previously Twitter) users and their peers. Across three large language models ranging in size from 1 billion to 70 billion parameters, we find that predicting a user's posts from their peers' posts performs poorly. Moreover, the value of the user's own posts for prediction is consistently higher than that of their peers'. Across the board, we find that the predictability of social media posts remains low, comparable to predicting financial news without context. We extend our investigation with a detailed analysis about the causes of unpredictability and the robustness of our findings. Specifically, we observe that a significant amount of predictive uncertainty comes from hashtags and @-mentions. Moreover, our results replicate if instead of prompting the model with additional context, we finetune on additional context.

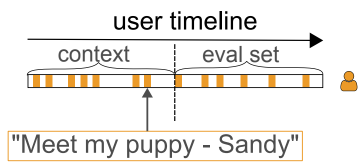
1 Introduction

Prediction is of fundamental importance for social research Salganik (2019); Salganik et al. (2020). The predictability of different social variables can provide a scientific window on a diverse set of topics, such as, emotion contagion Kramer et al. (2014) and social influence Bagrow et al. (2019); Cristali and Veitch (2022); Qiu et al. (2018), privacy concerns Garcia (2017); Garcia et al. (2018); Li et al. (2012); Jurgens et al. (2017), the behavior of individuals and groups Tyshchuk and Wallace (2018); Nwala et al. (2023), the heterogeneity of networks Colleoni et al. (2014); Aiello et al. (2012), information diffusion Chen et al. (2019); Weng et al. (2014); Bourigault et al. (2014); Guille and Hacid (2012) and more. Of particular importance for the study of digital platforms is the case of online speech. Studies about the predictability of text, however, have long been limited by the weaknesses of available language models.

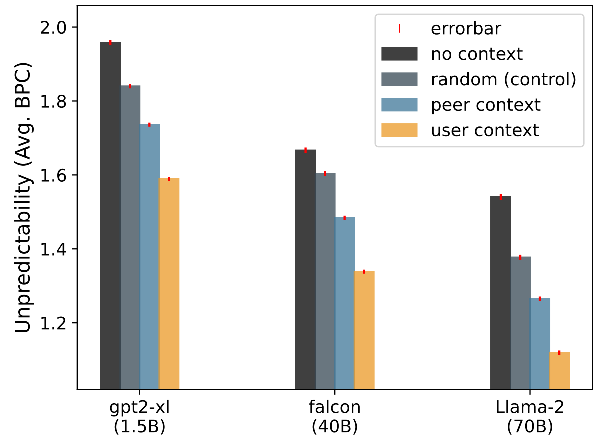
We revisit the problem of predicting online speech in light of dramatic advances language modeling. We focus on the central question: How predictable is our online speech using large language models? Such predictive capabilities could inform research on substantial risks – such as



(a) No context



(b) User context



(c) Average bits per character (BPC) required to predict user tweets, with 95% confidence intervals ($N = 5102$). Lower values mean greater predictability.

Figure 1: Predictability of a user's tweets using LLMs. Bits per character (BPC) measures, on average, how many bits are required to predict the next character. Predictability improves with additional context to the model: (i) past user tweets (user context, Fig. 1b) (ii) past tweets from the user's peers (peer context) and (iii) past tweets from random users (control). We plot the average BPC over users in Fig. 1c. **Most of the predictive information is found in the user context, followed by peer and random context.** Our results are robust across models with different parameter sizes and tokenizers.

user profiling, impersonation and exerting influence on real users Carroll et al. (2023); Weidinger et al. (2021). Inspired by work exploring the possibility of user profiling Bagrow et al. (2019), we explore the question of predicting user speech from their peers. We ask the following: How predictable is a social media post given posts from the author's peers? We contrast the answer with how predictable social media posts are from the authors' own posts. Through an experiment spanning millions of tweets and thousands of subjects, our study provides a detailed picture of the current state of predicting online speech.

1.1 Contributions

We investigate the predictability of online speech on the social media platform X (Twitter) using a corpus of 6.25M posts (tweets) of 5000 subjects and their peers. We use three large language models of increasing size, GPT-2-XL-1.5B, Falcon-40B, and Llama-2-70B. We use these models to estimate the predictability of our subjects' posts under various settings.

Predictability improves significantly with model size, as expected. However, **even the best model performs poorly without additional context**. For illustration, most users' posts are less predictable than financial news. We therefore test how much prediction benefits from three *additional* sources of context, specifically:

1. posts from randomly selected users (random context),
2. posts from the user's peers (peer context),
3. posts from the user themselves (user context).

Contrary to what prior work suggested Bagrow et al. (2019), our findings on state-of-the-art language models suggest that **a user's own posts have significantly more predictive information than posts from their close social ties**. Prediction benefits most from user context, followed by peer context, in turn followed by random context. We find that all of them improve predictability significantly, with a large effect size. Our prompting experiments in Figure 1 illustrate these findings.

To understand the sources of unpredictability, we further investigate the locations of greatest uncertainty within a post. We find that **@-mentions and hashtags are typically the points of greatest uncertainty** within a post. Consequently, removing both generally improves predictability across the board. The relative comparisons, however, stay unchanged.

Large language models can be sensitive to prompting strategies. To test the robustness of our findings, we replace prompting with fine-tuning. Specifically, instead of prompting GPT-2-XL with a specific source of context, we fine-tune the model on the same source of context. We find that the relative comparisons between user, peer, and random context, remain the same.

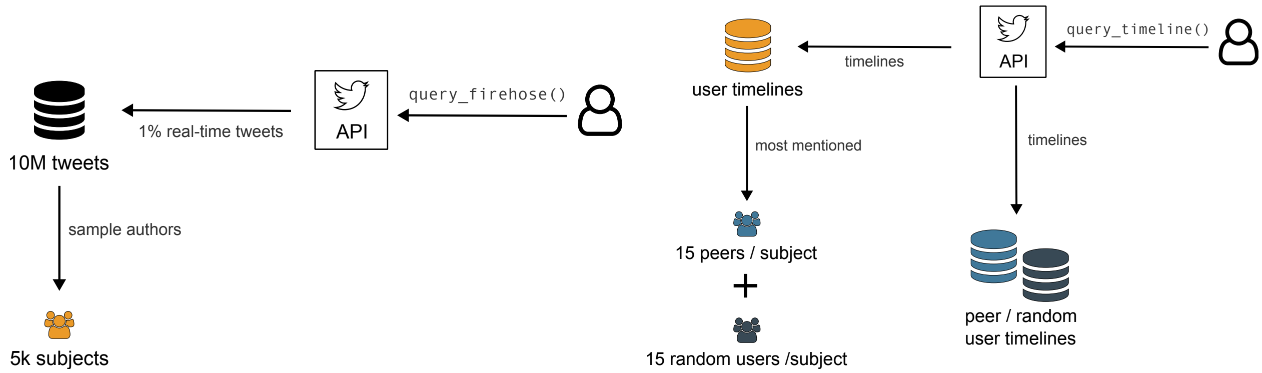
To summarize, the extent to which we can predict online speech is limited even with state-of-the-art language models. Our observations do not suggest that peers exert an outsize influence on an individual's online posts. Concerns that large language models have made our individual expression predictable are not supported by our findings.

2 Related work

Modeling online speech Understanding and modeling language usage on social media Bashlovkina et al. (2023); Kern et al. (2016); Schwartz et al. (2013) has become of particular interest to the research community, where Twitter is one of the most studied social media platforms Zhang et al. (2023); Qudar and Mago (2020). The most common method is using large language models (LLMs), where they finetune on social-media data and evaluate models on specific downstream tasks such as sentiment analysis, hate speech detection, etc. Zhang et al. (2023); Bashlovkina et al. (2023); Qudar and Mago (2020) We are interested in a more general notion of predictability – in particular, the predictability of individual users' online speech. Following common practice in natural language processing, we take model uncertainty as a measure of how hard it is for the model to predict some text.

Prediction based on neighbors Predicting attributes of a node from its neighbors is a well-known paradigm in machine learning. In the context of social networks however, it often has problematic implications on privacy since it limits the user's ability to control what can be inferred about them Garcia (2017); Garcia et al. (2018); Li et al. (2012). Work on the feasibility of "shadow profiles" (predicting attributes of non-users from platform users) have introduced a notion of privacy that is collective Garcia (2017); Garcia et al. (2018). Prediction of sensitive attributes of the user such as age, gender, religion etc. is possible from their peers Jurgens et al. (2017).

Bagrow et al. (2019) go beyond predicting sensitive attributes and look at the predictability of users' online speech. They investigate the theoretical possibility of peer-based user profiling on Twitter. They look at the information content of tweets using a non-parametric estimator, and derive an upper bound on predictability which shows that 8-9 peers suffice to match the predictive



(a) Sampling stage

(b) Timeline collection stage

Figure 2: Our data collection process can be divided into two stages. In the first stage (Fig. 2a), we collected 10M tweets in early 2023 which served as our base for sampling subjects. In the second stage (Fig. 2b), we collected users' timelines. First, we collected timelines of subjects. Based on those, we determined which users they had most frequently mentioned (their "peers"). We selected the top-15 most mentioned peers and 15 random users per subject, and collected their timelines too.

information contained in the user's own posts. This upper bound on predictability implies that there *could* exist some predictor which is capable of peer-based user profiling. Our work aims to contextualize their results by empirically testing this hypothesis on concrete predictors. We use transformer-based LLMs, which are currently considered to be the state-of-the-art method for language modeling.

3 Experimental setup

We consider a similar experimental setup as Bagrow et al. (2019), but with recent large language models and 15 peers per subject. We define predictability (or rather *un*predictability) as a measure of the model's uncertainty, i.e., its negative log-likelihood, on a specific user's tweets conditional on given context. We collected tweets from a sample of ∼5k subjects. We experiment with three types of contexts: (i) past tweets from the user (or "user context" - Fig. 1b), (ii) past tweets from the user's social circle ("peer context") and finally (iii) past tweets from a control group ("control" or "random context").

We start by describing our data collection process (Section 3.1) and what models we used (Section 3.2). Then, we go into detail about the implementation of our prompting and finetuning experiments (in Section 3.3 and 3.4, respectively).

3.1 Data Collection

The data collection process consisted of two main stages: an initial sampling period where we recorded real-time Twitter activity for a month in early 2023 and a second stage where we collected the timelines of sampled users. For a high-level overview of the data collection process, see Figure 2. We collected tweets from three groups of users:

1. *subjects*, approximately 5,000 randomly sampled Twitter users,

2. *peers* of subjects, which we take to be the top 15 people that each subject most frequently mentions,
3. *random* users for control purposes.

Throughout our data collection, we only collected tweets that were written by the user who posted them (e.g. no retweets) and were classified as English according to Twitter's own classification algorithm.

3.1.1 Sampling stage

Sampling was done by collecting a pool of tweets whose authors would serve as our base for picking our subjects. We collected them using the Twitter Firehose API; which allowed us to collect a 1% sub-sample of real-time tweet activity. This collection period lasted roughly 30 days and was done in early 2023 (from 20. January to 10. February), during which we collected 10M tweets (with $\sim$5M unique authors). We sampled $N = 5102$ subjects from this pool (0.1% of authors) for our experiments. Twitter had roughly $\sim$500M active monthly users in 2023 Yaccarino (2023) – some of which may have been bot accounts Lee et al. (2011). To address this issue, we decided to filter out accounts that scored high (above 0.5 on a scale of 0 to 1) on the Bot-O-Meter bot detection API Sayyadiharikandeh et al. (2020). Additionally, we also filtered out users that had a high retweet ratio, where $> 80\%$ of their feed consisted of retweets. This ensured the selection of users who had a sufficient amount of self-authored tweets for our dataset.

3.1.2 Timeline collection stage

For each subject, we collected a total of 500 tweets $\mathcal{T}_u = \mathcal{T}_u^{\texttt{eval}} \cup \mathcal{T}_u^{\texttt{user}}$ from their timelines. Half of those tweets were used for estimating predictability $\mathcal{T}_u^{\texttt{eval}}$, while the other half $\mathcal{T}_u^{\texttt{user}}$ served as context. Some tweets in $\mathcal{T}_u$ date back as early as 2011, however the vast majority (95%) were authored after 2022. Collected tweets were authored no later than the start of our sampling stage (20. January 2023).

Besides user context, we also introduce "peer" and "random" context: $\mathcal{T}_u^{\texttt{peer}}$ and $\mathcal{T}_u^{\texttt{random}}$, again with 250 tweets each. We selected the 250 most recent tweets from a pool of 15 users' timelines. Peers were the top-15 users who were most frequently mentioned by a subject. Random users were sampled from a total of $\sim$90k timelines that we had queried in this stage. We always made sure that context tweets were authored *before* the oldest tweet inside $\mathcal{T}_u^{\texttt{eval}}$. In total, we collected approximately 6.25M tweets for our experients.

Similarly to Bagrow et al. (2019), we created a second control group (*temporal control*). However, because of the similarity of the results, we only report results on the random control group (*social control*) in the main text. For results including the temporal control, please see Appendix B.4. For more detailed information on the dataset and how it was collected, we refer the reader to Appendix A.1. There we go into detail on what filters we used to achieve a high-quality, representative dataset of English tweets and share some statistics about our subject pool.

3.2 Models

We selected three large language models for our experiments: GPT-2-XL with 1.5B parameters Radford et al. (2019), Falcon with 40B parameters Almazrouei et al. (2023) and Llama-2 with 70B parameters Touvron et al. (2023). They represent models that share a similar transformer-based architecture that have recently become popular due to their generative capabilities. However, they differ in parameter size (1.5B, 40B and 70B parameters, respectively), training corpus and tokenizers. We considered using GPT-3 and GPT-4, however the OpenAI API unfortunately does not allow access to log probabilities for all tokens (which is necessary for our analysis), only to the top-5. We used Huggingface's transformers library Wolf et al. (2020) to load the models and run our experiments. For the prompting experiments, we loaded the larger models using 8-bit precision (which has little to no impact on performance Dettmers et al. (2022)) so that they could fit on 2 A-100 80GB GPUs. We only use GPT-2-XL for our finetuning experiments.

3.3 Prompting Experiments

Our first approach to measuring model uncertainty is through *prompting*; that is, experiments where we feed a tweet to a model and observe the associated probabilities of outputting that exact tweet. It is important to note that this approach does **not** involve content generation. This allows us to avoid a host of additional modeling choices, making prompting a more robust method. We use negative log-likelihood (or NLL) as a measure of model uncertainty, and introduce bits per character (or BPC) to enable comparisons across models. Reported results are calculated on $\mathcal{T}_u^{\texttt{eval}}$.

Let us denote a single tweet as $T = (t_1, t_2, ...t_m)$, where t_i is a single token ($i = 0...m$). A token is an item in the LLM's vocabulary, and can be thought of as a collection of characters that frequently co-occur. Each tweet has a maximum of 280 characters[1] and after tokenization most tweets have between 0-100 tokens (Fig. 14 in Appendix). We use a language model with parameters θ to predict token t_i based on the preceding tokens $t_{<i}$. The model's output will be a likelihood over all possible tokens in the model's vocabulary, however we are only interested in the conditional probability of t_i: $p_\theta(t_i|t_{<i})$.

3.3.1 Calculating negative log-likelihood

Our metric for predictability is the average negative log-likelihood (or NLL for short from now on). We define the NLL of a tweet T in the following way: $L(T) = -\sum_{t_i \in T} \ln p_\theta(t_i|t_{<i})$. This gives us an estimate of the model's uncertainty when predicting the tokens inside tweet T. The average uncertainty over all tweets in the evaluation set $\mathcal{T}_u^{\texttt{eval}}$ is

$$\bar{L}_u = \frac{1}{n} \sum_{T_j \in \mathcal{T}_u^{\texttt{eval}}} L(T_j),$$

where n is the total number of tokens. We additionally introduce notation to distinguish what context was used to calculate: $\bar{L}_u^c$, where subscript $c \in \{\texttt{user}, \texttt{peer}, \texttt{random}\}$. Here, the conditional probability of token t_i is based on preceding tokens $t_{<i}$ as well as tokens from the appropriate context: $p_\theta(t_i|\mathcal{T}_u^c, t_{<i})$. The added context lends a convenient cross-entropy like interpretation of the shared information between the context and evaluation tweets we are trying to predict. Tweets

[1]This limit changed to 4000 characters on the 09. February 2023.

that served as context were concatenated using the 'newline' token for GPT-2-XL and the 'space' token for the two bigger models (they had no standalone 'newline' token like GPT-2). We used the maximum input sequence length available for each model (1024 tokens for GPT-2-XL, 2048 for Falcon 40B and 4096 for Llama-2) to calculate the token probabilities. In case the provided context exceeded this length, the oldest tweets were discarded.

3.3.2 Conversion to bits per character

Models with different tokenizers produce NLLs that are not commensurate with each other due to a different set of tokens being used. To overcome this, we convert our measure to a metric called *bits per character*, also known as *bits per byte*. Let c_t be the number of characters in token t. We define the number of characters inside tweet T as $C(T) = \sum_{t_i \in T} c_{t_i}$. Taking the average over all tokens in $\mathcal{T}_u$ we get $\bar{C}_u = \frac{1}{n} \sum_{T_j \in \mathcal{T}_u} C(T_j)$. We define bits per character (BPC) formally as

$$\overline{bpc}_u = \bar{L}_u \cdot \frac{1}{\bar{C}_u} \cdot \frac{1}{\ln 2}\,.$$

This number tells us the average number of bits required to predict the next character in the set $\mathcal{T}_u^{\mathtt{eval}}$. This leads to a more interpretable variant of the common perplexity measure [2]. The average BPC over all users is $\overline{bpc} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \overline{bpc}_u$. Similarly to before, we define $\overline{bpc}^c$ as the measure of model uncertainty where token probabilities were calculated including tokens from some context c.

3.3.3 Estimating the effect size of context on predictability

Next, we are interested in measuring the average effect size of different contexts on predictability. At a user level, we quantify the difference in predictability between context $c1$ and $c2$ as $\Delta_{c2}^{c1}(u) = \bar{L}_u^{c1} - \bar{L}_u^{c2}$. Aggregating over all users, we define the standardized mean difference (SMD) of Δ_{c2}^{c1} for each context pair:

$$\mathtt{SMD}(\Delta_{\mathtt{c2}}^{\mathtt{c1}}) = \frac{\mu_{\Delta_{\mathtt{c2}}^{\mathtt{c1}}}}{\sigma_{\Delta_{\mathtt{c2}}^{\mathtt{c1}}}}$$

This gives us a unit-free quantity to estimate the effect size. Following established practice Cohen (1988), we characterize effect sizes up to 0.2σ to be small, up to 0.8σ to be medium and and anything above that to be a large effect size.

3.4 Finetuning Experiments

In this section we present a second method for estimating predictability, which has two advantages over our previous prompting approach. First, we can fit the entire context into the model compared to prompting where the maximum size of the input sequence was a limitation. Second, it bypasses a common ailment of LLMs: their sensitivity to prompting strategy. However, finetuning can also be very expensive—which is why we decided to limit our experiments to GPT-2-XL only.

Tweets were concatenated (with the special eos token '<|endoftext|>' serving as a separator between them) then divided into chunks of 1024 tokens so that they fit into the model. Here,

[2]Perplexity is the exponentiated average negative log-likelihood.

our proxy for unpredictability was the negative log-likelihood (or also commonly known as the *cross-entropy* loss) in the final round of fine-tuning. Reported results are calculated on $\mathcal{T}_u^{\texttt{eval}}$.

Our finetuning experiments had two stages. In the first *pre-finetuning* stage, we trained on a large set of tweets to condition the model to learn basic Twitter syntax and common vocabulary. This later helped us learn user related features much more efficiently in the second stage. After pre-finetuning, we *additionally finetuned* on one of the contexts. We also experimented with using *mixtures* of different contexts. We used the example finetuning script from the transformers library as our base[3], where we kept most of the default training arguments. Some finetuning choices were inspired by an open-source project called HuggingTweets Dayama (2022).

3.4.1 Pre-finetuning on 10M tweets

We finetuned GPT-2-XL on the 10M tweets we collected during the sampling stage of our data collection process (as described Section 3.1), where a 5% split was reserved for validation. We finetuned all parameters for 1 epoch, using a constant learning rate of 5e$-$5, batch size 8 and `fp16` mixed precision training on a single A-100 80GB GPU. More details on the pre-finetuning and loss curves can be found in Appendix C.1.

3.4.2 Finetuning on context

Starting from the pre-finetuned model, for each user we finetuned a model on $\mathcal{T}_u^c$, for each $c \in \{\texttt{user}, \texttt{peer}, \texttt{random}\}$. That is three separate models per subject, resulting in a total of $\sim$15k finetuned models. We finetuned for 5 epochs with constant learning rate 1e$-$5 and batch size 1. We tracked the cross-entropy loss on $\mathcal{T}_u^{\texttt{eval}}$ periodically.

We also include experiments on finetuning on a mixture of contexts (eg. `peer+random` containing tweets from both $\mathcal{T}_u^{\texttt{peer}}$ and $\mathcal{T}_u^{\texttt{random}}$). We combined them by sampling an equal amount of tweets uniformly from each context, while keeping the total number of tweets the same (250 tweets) to enable a fair comparison. Again, the base model was our pre-finetuned model, and we finetuned for 5 epochs with constant learning rate 1e$-$5 and batch size 1.

4 Results

We divide the presentation of our main findings into three parts, one about the unpredictability of posts in general (Section 4.1), one about predicting from different sources of context (Section 4.2), and one about our investigation into the sources of unpredictability (Section 4.3).

4.1 Unpredictability of Posts

As described in Section 3.3, we used negative log-likelihood as a measure of the model's uncertainty when predicting a user's tweets. In Figure 3 we present the distribution of the average NLL of our subjects. To gain an intuition on how hard it is to predict our subjects' tweets relative to other accounts, we included three popular financial news accounts as comparison: YahooFinance,

[3] `run_clm.py` from here: https://github.com/huggingface/transformers/tree/main/examples/pytorch/language-modeling

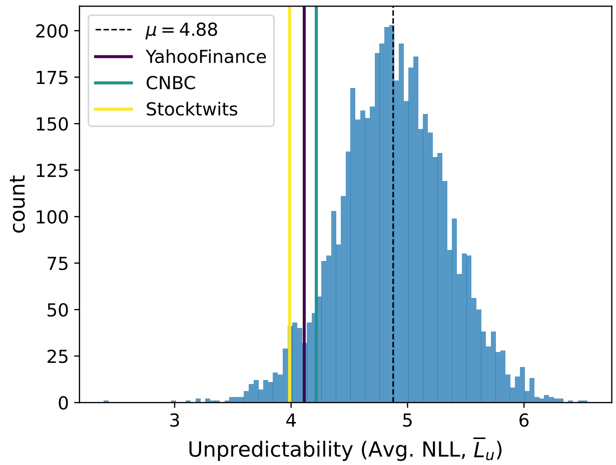


Figure 3: Distribution of the average NLL of our subjects (with no additional context). As comparison we include the average NLL of three financial news accounts: YahooFinance, CNBC and Stocktwits. Colloquially financial news are hard to predict, therefore most of our subjects are too. Model: GPT-2-XL.

CNBC, and Stocktwits. Intuitively, predicting financial news and the stock market is hard Johnson et al. (2003), which should make them less predictable. However, Figure 3 reveals that most subjects in our pool are actually *harder* to predict than those news accounts. Of course, while the *content* of those tweets may be hard to predict, their *style* and *vocabulary* may not.

We find that the ranking of users according to average unpredictability of their tweets is strongly correlated between all three of our models (Figure 4). This means that a user who scores high (i.e. their tweets are harder to predict) according to one model, will likely score high on a different model as well. This points to a certain robustness of the negative log-likelihoods: Although the absolute numbers change from one model to the next, the ranking of users is similar.

Finally, the absolute information content of users' tweets (without additional context) is between 1.5-2 bits per character, depending on which model we use (see 'no context' bars in Figure 1c). As a comparison, Shannon's upper bound on the cross-entropy of the English language is 1.3 bits per character Shannon (1951), and recent estimates using neural language models say it is as low as 1.12 bits Takahashi and Tanaka-Ishii (2018). Only our largest model with additional user context achieves comparable entropy, with an average of 1.1193 bits per character (Llama-2 with user context). These results suggest that individual pieces of our online expression taken out of context are far from predictable using today's LLMs.

4.2 Predicting Posts from Context

We quantified how additional context influenced predictability. We used two methods of feeding additional information to our model, with complementary strengths and weaknesses: one prompting based approach and one finetuning based approach.

We start with the results of the prompting approach. In Figure 1, we plot the average BPC calculated over all of our subjects for three different models with varying parameter sizes. The no context case was consistently the most unpredictable with highest BPC. Depending on what

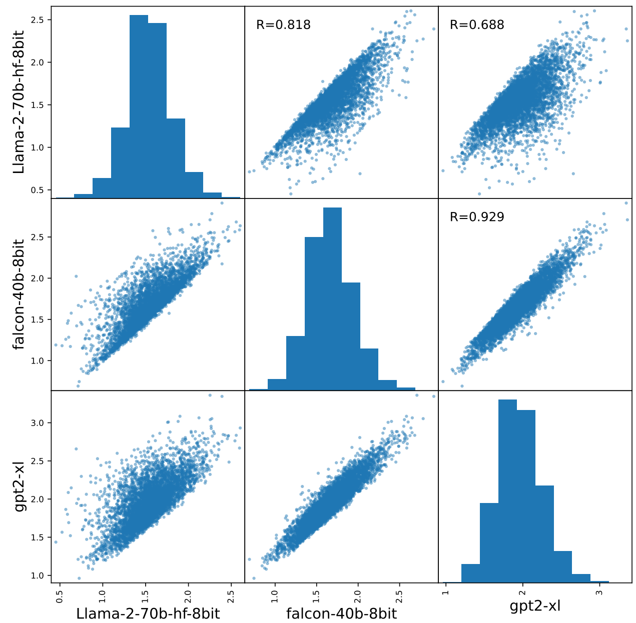
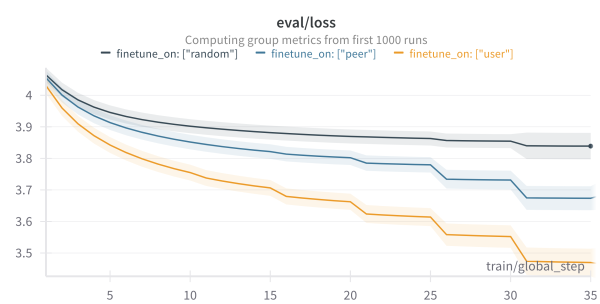
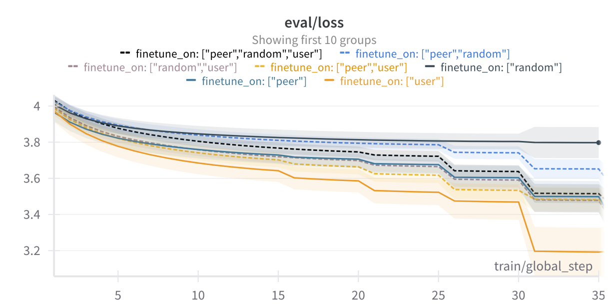


Figure 4: There is high agreement on the ranking of users based on predictability across models. We measure predictability in the no-context setting ($\overline{bpc}_u$; bits per character) for each user. A high-scoring user (who is harder to predict) will get a similarly high score on a different model (conversely a low-scoring user will get a lower score).



(a) Finetuning on user, peer and random context exclusively. Again, most predictive information can be found in the user context, followed by peer then random context.



(b) Finetuning on combinations of contexts. For example ["peer", "random"] denotes that half of the tweets were sampled from $\mathcal{T}_u^{\texttt{peer}}$ and the other half from $\mathcal{T}_u^{\texttt{random}}$. Mixing contexts does not always lead to better results, suggesting overlap in predictive information.

Figure 5: Average loss curves with standard error for finetuning experiments on GPT-2-XL. For each subject ($N = 5102$) we finetune on the specified context, and compute the loss (negative log-likelihood) on $\mathcal{T}_u^{\texttt{eval}}$. The plotted averages are computed over the loss curves of 1000 subjects.

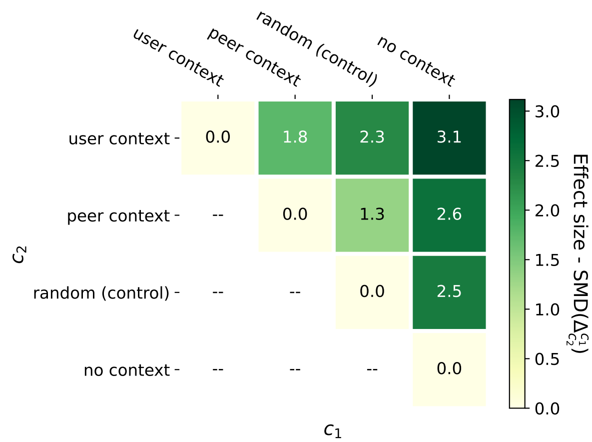


Figure 6: Effect size matrix on Llama-2. We plot the average effect size of c_2 relative to c_1 on user predictability. This is done by looking at the difference in predictability for each user $\Delta_{c2}^{c1}(u) = \bar{L}_u^{c1} - \bar{L}_u^{c2}$, where $\bar{L}_u^c$ is the average negative log-likelihood of user u under context c. Plotted values are standardized mean differences (SMD) of Δ_{c2}^{c1}. Darker green means greater improvement in prediction.

context was included, the amount of improvement varied. We show that most of the predictive information is found in the user context, followed by peer and random context. This trend is consistent across all three of our models.

As part of our prompting experiments, we were also interested in quantifying the effect size each context had on predictability. We plot this for each context pair in Figure 6 on Llama-2. This effect size matrix illustrates our finding that all types of context have a large effect size ($> 2.5\sigma$) over having no context at all (last column). It also allows us to quantify the effect size of our main finding: user context offers 1.8σ improvement over peer context — an even bigger improvement than what peer context offers over random context (1.3σ). While these effect sizes are most pronounced on Llama-2, we found that these results also translate to the other models as well (Figure 17 in Appendix).

Our main finding was replicated in our finetuning experiments as well. Figure 5a shows the average loss curves (computed over 1000 subjects) for finetuning GPT-2-XL on different types of contexts. Again, for all three contexts the loss goes down significantly. However, the final loss they converge to is different, with large gaps between each. If we order contexts based on the achieved loss in the final round, we get the same order as before: user context is best, followed by peer context, then by the random control context. Bagrow et al. (2019)'s findings suggest that with enough peers, one can match the predictive information inside the user context (using 8-9 peers). However even with 15 peers per subject, we find no evidence of such peer-based user profiling.

We established that user context always outperfors peer context, regardless of model choice or experimental method. Still, one might argue that by *combining* user and peer context, one might achieve even higher predictability than by using user context alone. This could suggest that there is some non-overlapping predictive information inside the peer context wrt. the user context. We test this hypothesis by finetuning on a mixture of contexts (Fig. 5b) to give us an idea which contexts shared common information. Combining `peer+random` contexts resulted in

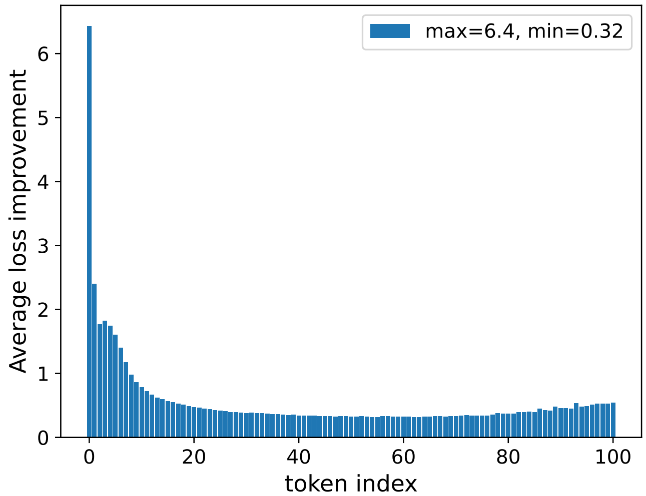


Figure 7: Average improvement in NLL from additional user context (compared to none). The first few tokens of a tweet benefit most from the additional context. Model: Llama-2. We plot up to 100. token, because the vast majority of tweets have < 100 tokens (Fig 13 in Appendix).

a linear interpolation of the final losses of finetuning on either context. In other words, mixing random and peer context outperformed the final loss of finetuning exclusively on random context by a large margin. Here, we found evidence of additional predictive information in the peer context, which was not contained in the random context. However, mixing `peer+user` contexts did *not* result in a significantly lower final loss. This suggests that there is a significant overlap in predictive information between peer and user context.

A final, clarifying note on the "jumps" in the loss curves: Since each $\mathcal{T}_u^c$ contains 250 tweets of varying lengths, 5 epochs of training resulted in different global steps for each user-context combination. This explains the irregularities around global steps 15, 20, ... etc. since for some users we only have losses up to that global step.

4.3 What Drives Unpredictability

Finally, we were curious where and how context improved predictability. Perhaps a mildly surprising observation we can make based on Figure 1 is that random context also improves predictability in a non-negligible way. Due to the randomized nature of the control group there should be very little connection to our subjects' tweets *semantically*. However the improvement we observed can be explained by added information regarding the *syntax* of a tweet – how long it is, the presence of @-mentions and hashtags, etc. Indeed this turns out to be the case: with random context, the predictability of the @ token improved the most compared all other tokens – see Appendix Table 1.

Even more surprisingly, predicting @-mentions and hashtags correctly played a significant role even in the case of user and peer context. Visualizing the improvement in predictability over individual tokens inside a tweet we noticed an interesting phenomenon (Figure 7). The

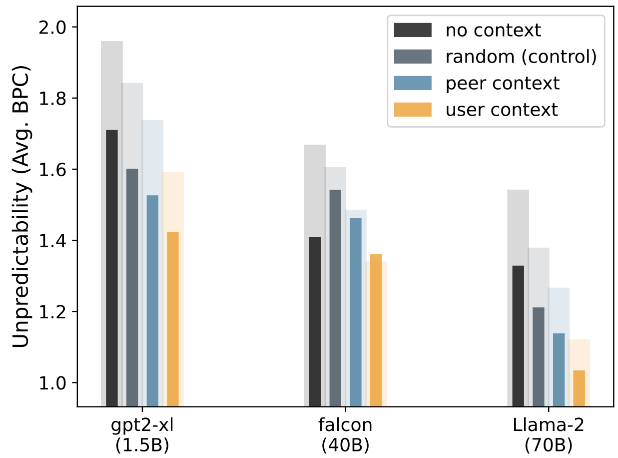


Figure 8: Rerunning our experiments without @-mentions and hashtags results in these changed predictabilities. Subjects become more predictable on average, and the positive effect of context on predictability decreases. For Falcon it almost completely vanishes (with only user context offering slight improvement). Lighter bars are the results from our original experiment for comparison.

initial"spike" in the figure suggests that some of the most improved tokens were at the start of tweets, often tokens part of @-mentions. After that, the average improvement stabilizes, then slightly goes up towards the end of the tweets, where hashtags would often be located. Based on these results it seemed plausible that @-mentions and hashtags were locations of greatest uncertainty, where additional context helped model prediction the most.

We repeated our prompting experiments after removing @-mentions and hashtags from our tweets. Assuming these were indeed one of the greatest sources of error, removing them would make our tweets more predictable overall, and the effect of additional context on predictability less strong. The results of our repeated experiments can be seen in Figure 8. Indeed, subjects became more predictable on average, and the effect of context on predictability decreased (by $\sim 20\%$ for random context, $\sim 4\%$ for peer context and $\sim 7\%$ for user context on Llama-2 70b). One peculiar finding was that for Falcon additional context did not improve predictability the same way as before. For example, the effect size of the user context dropped to only 0.4σ, which is significantly less than the 2.1σ from before. A possible explanation is that the only (useful) predictive signal Falcon was able to pick up on was in the removed pieces of text. Another possibility is that it is simply more sensitive to discontinuities inside the text than other models.

5 Limitations

Negative-log likelihood as a measure of uncertainty While negative log-likelihood is the most popular measure of model uncertainty, there are certain limitations to using this metric. One is that it is calculated exclusively on tokens contained in tweet T, and does not take into account any improvements outside of that token (i.e. other tokens that are close to it in the embedding space). Another issue is the sensitivity of this metric: to the prediction of the first token (see Fig. 7 and Appendix B.5) and to slight changes in prompting (how tweets were separated, for example). Using low-frequency tokens as a separator between tweets (such as the eos token, which is usually reserved for separating training documents) produced abnormally high NLLs, which is why we decided to use more common tokens, such as space or newline for the prompting experiments. Another important aspect is the length of input; NLL is commonly lower for longer pieces of text. This is true here as well, and we show the relationship between average tweet length and model uncertainty in Appendix B.2. While these points affect absolute numbers in our analysis, it does not affect our main statements about the relative comparisons between different models and contexts. As an example, our choice of model also had a considerable effect on the absolute value of BPC (with GPT-2-XL scoring typically higher than Llama-2), however it did not affect the *relative* relationship between the different settings.

Data contamination Furthermore, we cannot rule out the possibility that some tweets in our corpus were part of the training dataset of the LLMs that we used. However, we believe the risk and severity of data contamination to be limited. GPT-2 was trained on WebText with content up to December 2017 Radford et al. (2019) (95% of our subjects' tweets were authored after 2022), Falcon was trained on RefinedWeb which is is built using all CommonCrawl dumps until the 2023-06 one Penedo et al. (2023) (CommonCrawl typically does not contain snapshots of Twitter com (2024)) and the authors of Llama-2 claim to have "made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals." Touvron et al. (2023).

6 Discussion

We presented the results of an investigation using state-of-the-art large language models into the predictability of online speech by analyzing posts on X (Twitter). As the basis of our study, we collected posts from more than five thousand users' timelines and their peers. We used a total of 6.25M tweets for our prompting experiments, plus an additional 10M for finetuning.

We found that it is surprisingly difficult to predict online speech. The majority of our users' tweets are less predictable than tweets about financial news. Even with additional context, prediction performs relatively poorly. Only our largest model with additional user context is able to approximate the estimated information content of the English language. Further analysis on what drives this unpredictability reveals that the source of greatest uncertainty are hashtags and @-mentions. In other words, what improves prediction the most with additional context are basic signals: specifically, guessing the correct hashtags and @-mentions. All in all, our findings suggest that despite the impressive capabilities of large language models in other areas, state-of-the-art models predict online speech rather poorly.

Another central question of our analysis was whether the predictive information inside peer tweets is enough to match (or even surpass) those of the user's own tweets, as suggested by Bagrow et al. (2019). We believe that this is unlikely. Our results show that user context consistently outperforms peer context in a manner that is robust to model choice and evaluation method. Additional experiments suggest that whatever predictive information is inside the peer context, can also be found in the user context.

Our findings may also inform work on privacy and other ethical issues on social networks. For example, threats such as shadow profiling or impersonation on a global scale may not be as acute as some feared with state-of-the-art language models. However, these risks may still exist on a case-by-case basis (we demonstrate individual variability in Figure 12 in the Appendix). Moreover, there may be other conceivable harms that don't map cleanly to questions of predictability.

7 Future Work

We presented results on the predictability of english speaking users' online speech. This is only the first step towards contextualizing what it means for individuals to live in a world of transformers-based language models. It still remains to be answered how these results might translate to different sub-communities. Extending our analysis to different sub-communities and languages would be an interesting avenue for future work. Similarly, we welcome additional work into the external validity of our results. It is unclear to what extent our findings would translate to other social media platforms.

Finally, we find that additional research into the downstream effects of predictable speech on privacy is necessary. One well known issue about LLMs is that they are known to memorize personally identifiable information (PII) Lukas et al. (2023). We are happy to see a shift in recent literature acknowledging that preventing data leakage is not equivalent to meaningful privacy guarantees in the natural language context Brown et al. (2022).

Acknowledgements

We are grateful to Ana Andrea Stoica, Ricardo Dominguez and Vivian Nastl for their valuable feedback on the initial drafts of this project. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Mina Remeli. This work was supported by the Tübingen AI Center.

References

Statistics of Common Crawl Monthly Archives by commoncrawl, 2024. URL https://commoncrawl.github.io/cc-crawl-statistics/plots/domains.

Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2):1–33, 2012.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon Series of Open Language Models, November 2023. URL http://arxiv.org/abs/2311.16867. arXiv:2311.16867 [cs].

James P. Bagrow, Xipei Liu, and Lewis Mitchell. Information flow reveals prediction limits in online social activity. *Nature Human Behaviour*, 3(2):122–128, February 2019. ISSN 2397-3374. doi: 10.1038/s41562-018-0510-5. URL http://arxiv.org/abs/1708.04575. arXiv:1708.04575 [physics].

Vasilisa Bashlovkina, Riley Matthews, Zhaobin Kuang, Simon Baumgartner, and Michael Bendersky. SMILE: Evaluation and Domain Adaptation for Social Media Language Understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pages 3737–3749, New York, NY, USA, August 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599907. URL https://dl.acm.org/doi/10.1145/3580305.3599907.

Simon Bourigault, Cedric Lagnier, Sylvain Lamprier, Ludovic Denoyer, and Patrick Gallinari. Learning social network embeddings for predicting information diffusion. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 393–402, 2014.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292, 2022.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, 2023.

Xueqin Chen, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Fengli Zhang. Information diffusion prediction via recurrent cascades convolution. In *2019 IEEE 35th international conference on data engineering (ICDE)*, pages 770–781. IEEE, 2019.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, New York, 2 edition, July 1988. ISBN 978-0-203-77158-7. doi: 10.4324/9780203771587.

Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332, 2014.

Irina Cristali and Victor Veitch. Using embeddings for causal estimation of peer influence in social networks. *Advances in Neural Information Processing Systems*, 35:15616–15628, 2022.

Boris Dayama. HuggingTweets: Train a Model to Generate Tweets, 2022. URL https://wandb.ai/wandb/huggingtweets/reports/Hugging-Tweets-Train-a-Model-to-Generate-Tweets--VmlldzoxMTY5MjI.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale, November 2022. URL http://arxiv.org/abs/2208.07339. arXiv:2208.07339 [cs].

David Garcia. Leaking privacy and shadow profiles in online social networks. *Science Advances*, 3(8):e1701172, August 2017. doi: 10.1126/sciadv.1701172. URL https://www.science.org/doi/full/10.1126/sciadv.1701172. Publisher: American Association for the Advancement of Science.

David Garcia, Mansi Goel, Amod Kant Agrawal, and Ponnurangam Kumaraguru. Collective aspects of privacy in the Twitter social network. *EPJ Data Science*, 7(1):1–13, December 2018. ISSN 2193-1127. doi: 10.1140/epjds/s13688-018-0130-3. URL https://

epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0130-3. Number: 1 Publisher: SpringerOpen.

Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. Of Bots and Humans (on Twitter). In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 349–354, Sydney Australia, July 2017. ACM. ISBN 978-1-4503-4993-2. doi: 10.1145/3110025.3110090. URL https://dl.acm.org/doi/10.1145/3110025.3110090.

Adrien Guille and Hakim Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 1145–1152, 2012.

Neil F. Johnson, Paul Jefferies, and Pak Ming Hui. *Financial Market Complexity*. Oxford University Press, July 2003. ISBN 978-0-19-171210-4. doi: 10.1093/acprof:oso/9780198526650.001.0001. URL https://academic.oup.com/book/9066.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Writer Profiling Without the Writer's Text. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, editors, *Social Informatics*, volume 10540, pages 537–558. Springer International Publishing, Cham, 2017. ISBN 978-3-319-67255-7 978-3-319-67256-4. doi: 10.1007/978-3-319-67256-4_43. URL http://link.springer.com/10.1007/978-3-319-67256-4_43. Series Title: Lecture Notes in Computer Science.

Margaret L Kern, Gregory Park, Johannes C Eichstaedt, H Andrew Schwartz, Maarten Sap, Laura K Smith, and Lyle H Ungar. Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, 21(4):507, 2016.

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, June 2014. doi: 10.1073/pnas.1320040111. URL https://www.pnas.org/doi/abs/10.1073/pnas.1320040111. Publisher: Proceedings of the National Academy of Sciences.

Kyumin Lee, Brian Eoff, and James Caverlee. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):185–192, 2011. ISSN 2334-0770. doi: 10.1609/icwsm.v5i1.14106. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14106. Number: 1.

Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1023–1031, New York, NY, USA, August 2012. Association for Computing Machinery. ISBN 978-1-4503-1462-6. doi: 10.1145/2339530.2339692. URL https://doi.org/10.1145/2339530.2339692.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing Leakage of Personally Identifiable Information in Language Models, April 2023. URL http://arxiv.org/abs/2302.00539. arXiv:2302.00539 [cs].

Alexander C Nwala, Alessandro Flammini, and Filippo Menczer. A language framework for modeling social media account behavior. *EPJ Data Science*, 12(1):33, 2023.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only, June 2023. URL http://arxiv.org/abs/2306.01116. arXiv:2306.01116 [cs].

Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2110–2119, 2018.

Mohiuddin Md Abdul Qudar and Vijay Mago. Tweetbert: a pretrained language representation model for twitter text analysis. *arXiv preprint arXiv:2010.11091*, 2020.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Matthew J Salganik. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.

Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 2725–2732, New York, NY, USA, October 2020. Association for Computing Machinery. ISBN 978-1-4503-6859-9. doi: 10.1145/3340531.3412698. URL https://doi.org/10.1145/3340531.3412698.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

C. E. Shannon. Prediction and entropy of printed English. *The Bell System Technical Journal*, 30 (1):50–64, January 1951. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1951.tb01366.x. URL https://ieeexplore.ieee.org/document/6773263. Conference Name: The Bell System Technical Journal.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. Cross Entropy of Neural Language Models at Infinity—A New Bound of the Entropy Rate. *Entropy*, 20(11):839, November 2018. ISSN 1099-4300. doi: 10.3390/e20110839. URL https://www.mdpi.com/1099-4300/20/11/839. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel

Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Yulia Tyshchuk and William A Wallace. Modeling human behavior on social media in response to significant events. *IEEE Transactions on Computational Social Systems*, 5(2):444–457, 2018.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Predicting successful memes using network and community structure. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 535–544, 2014.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing, October 2020. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6. Pages: 38–45 original-date: 2018-10-29T13:56:00Z.

Linda Yaccarino. One year in, the future of x is bright, 2023. URL https://blog.twitter.com/en_us/topics/company/2023/one-year-in.

Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and Generalizable Social Bot Detection through Data Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1096–1103, April 2020. ISSN 2374-3468. doi: 10.1609/aaai.v34i01.5460. URL https://ojs.aaai.org/index.php/AAAI/article/view/5460. Number: 01.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5597–5607, 2023.

A Experimental Setup

A.1 Data collection

A.1.1 Sampling stage

We used the Twitter Firehose API to query a 1% sub-sample of real time tweet activity in early 2023 (from 20. January to 10. February). We used the following filter expression to query the API:

```
'sample:1 followers_count:0 -is:retweet lang:en'
```

Where the options have following meaning:

- `sample:1`: Return a 1% sub-sample of the filtered tweets. (The specified number has to be between 1-100 representing a % value.)
- `followers_count:0`: Return tweets made by users with at least 0 number of followers. This is a dummy filter because sample/is/lang are not standalone filters (filters that can be used on their own) and need additional standalone filters (like followers_count) to work.
- `-is:retweet`: Don't return retweets.
- `lang:en`: Return English tweets.

This collection phase resulted in a pool of 10M tweets, with 5M unique authors. We sampled our subjects from this pool of authors ($N = 5102$ sample). Strictly speaking, our sample will be biased towards users that have 1) been more active during our initial collection period and 2) are more active users in general. This was the closest we could get to a random sample with the offered API endpoints. It is also follows Bagrow et al. (2019)'s method of randomly sampling users. Again, following in their steps we dropped subjects that had scored high using the Bot-O-Meter API. Bot-O-Meter uses classifies users on a scale from 0-1 based on 200 of their tweets. See Figure 9 for a distribution of these scores. We dropped users that had a score higher than 0.5. Additionally, we dropped users that had a high retweet ratio (more than 80% of their tweets consisted of retweets). This is an additional measure to prevent bot accounts in our subject pool (bots are known for frequently retweeting content Yang et al. (2020); Gilani et al. (2017)) as well as a practical consideration since we only wanted to include non-retweets in our analysis.

A.1.2 Timeline collection stage

We used the Twitter API's Timeline endpoint to query the subjects' most recent tweets. To get tweets from around the same timeframe, we choose an `end_time` (which was the start of our sampling stage, 20. January). Tweets made after `end_time` were not included. Per subject, we collected a total of 500 tweets, half of which was reserved for evaluation, the other half served as user context. From these tweets, we identified the top-15 most mentioned users (the user's peers) and collected 50 tweets from their timelines as well. These, as well as the user context tweets were collected such that they were authored before the oldest evaluation tweet (t^* in Figure 10). From these, we selected the 250 most recent tweets. A simpler approach would have been to select exactly $\frac{250}{15} \approx 17$ tweets per peer. However, not all peers had this many tweets on their timeline before t^*, hence our choice for collecting more tweets per peer than necessary.

In the end, we had queried the timelines of around ∼90k users, with 15M timeline tweets in our database. We sampled from this pool of users and tweets for our social / temporal control. The social control consists of tweets of 15 random users with $\frac{250}{15} \approx 17$ tweets each. We made sure that the sampled user did not coincide with the subject itself / any of their peers. The temporal control on the other hand contains tweets that were made around the same time as the tweets inside peer control. Again, we made sure that the tweets' authors did not overlap with the peer pool or the subject. Figure 10 illustrates all of our settings, while Figure 11 shows the time histogram of an example dataset belonging to one of our subjects.

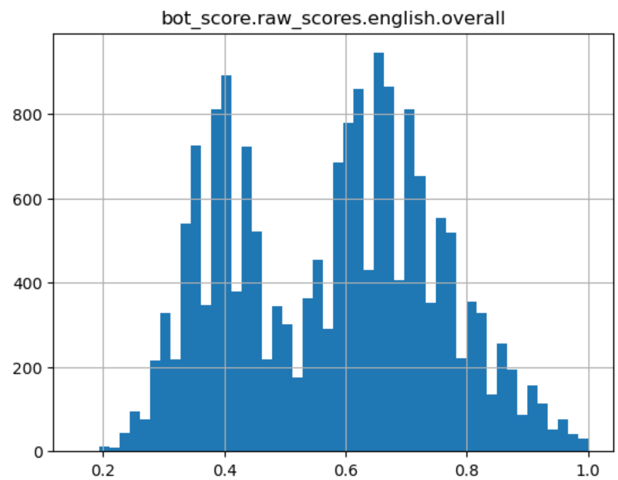


Figure 9: Distribution of Bot-O-Meter scores. We only selected users in our subject pool that had a score that was lower than 0.5.

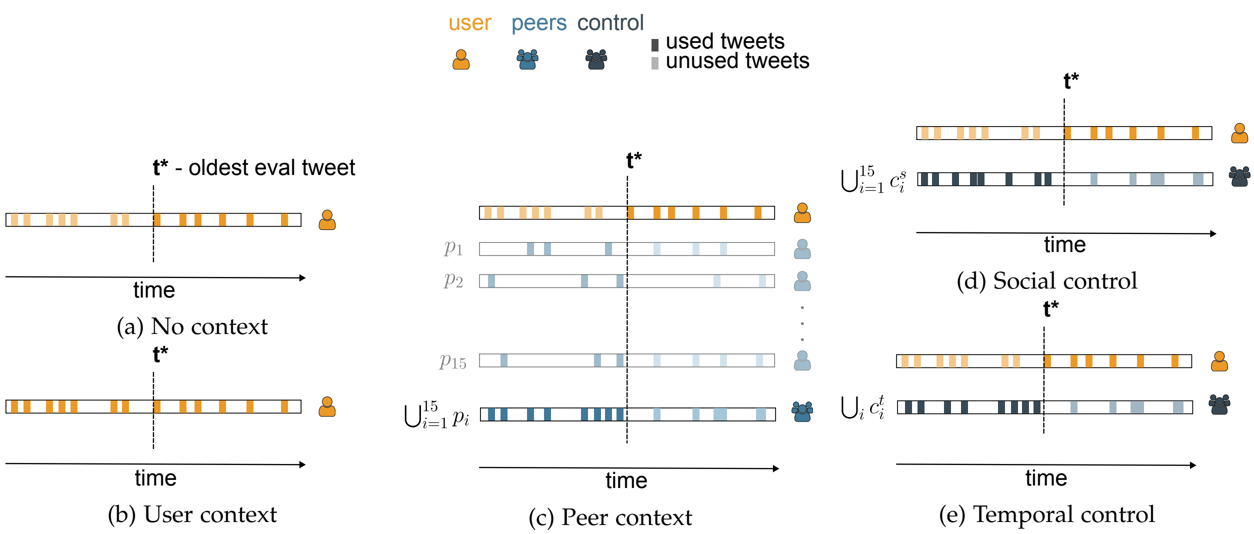


Figure 10: Data creation protocol. We evaluate predictability on a set of evaluation tweets, and how it changes depending on what context we provide.

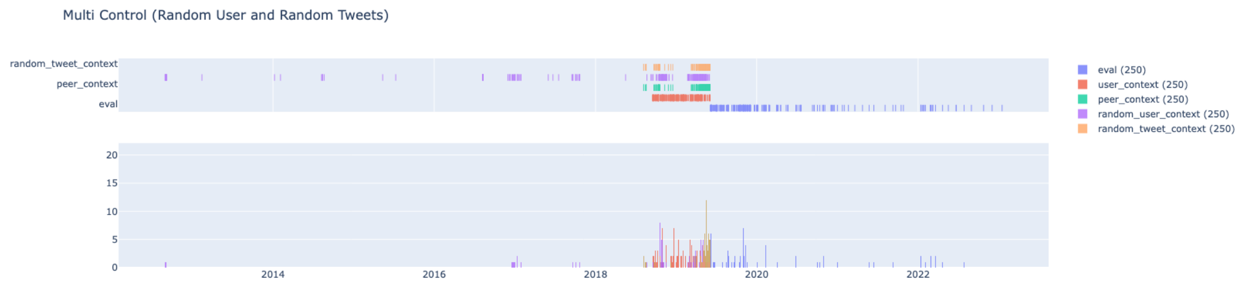


Figure 11: Time histogram of an example dataset of some subject u. We illustrate $\mathcal{T}_u^{\texttt{eval}}$, and how all context tweets were written before the oldest tweet in $\mathcal{T}_u^{\texttt{eval}}$. The social control contains tweets from random users, while the temporal control contains tweets that were authored around the same time as the peer tweets.

A.1.3 Preprocessing tweets

Before our experiments, we preprocessed our collected tweets by filtering out urls, deduplicated spaces and fixed some special character encodings. We found that urls were not relevant in analysing the predictability of (organic) online speech, while deduplication of spaces is a fairly common preprocessing step in NLP. For our experiments described in Section 4.3, we further filtered out @-mentions and hashtags.

B Prompting

B.1 Individual variability

While globally there is a tendency where user context outperforms peer context, peer context outperforms random context, etc., there is substantial variability on the individual level as illustrated in Figure 12. In the highlighted blue example, random context improves predictability more than the peer context.

B.2 Correlation with average tweet length

In Figure 13 we show the distribution of subjects' tweet length in tokens. Most of the tweets have between 0 and 100 tokens. We also present the relationship between model uncertainty and the average tweet length of a given subject in Figure 14. NLL and average tweet length are correlated: users with longer tweets are on average more predictable.

B.3 Hard to predict users gain more from additional context

In Figure 4 we have shown that if we rank users according to how predictable they are, there is a high agreement across models wrt. this ranking. Now, with additional context, we analyze how the change in predictability is influenced by this ranking (Figure 15). We find that with additional context, improvement in predictability is larger for users who are already hard to predict. Predictability improves on average by ~ 0.1 bits for every 1 bit increase in "difficulty to predict".

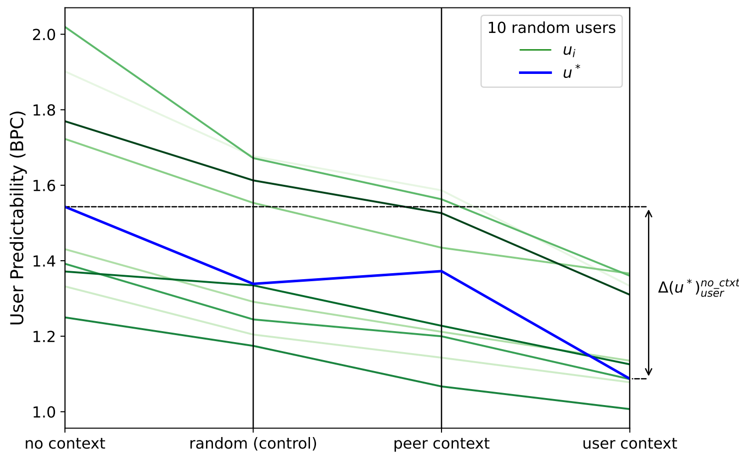


Figure 12: Differences in predictability $\Delta^{c_1}_{c_2}$ using the Llama-2 model. We picked 10 random users and plot their predictability using different contexts (y axis). Comparing across contexts we get the difference in predictability: $\Delta(u_i)^{c_1}_{c_2}$. While unpredictability goes down on average as we evaluate on more predictive contexts, there can be individual level variance that does not necessarily always follow this trend.

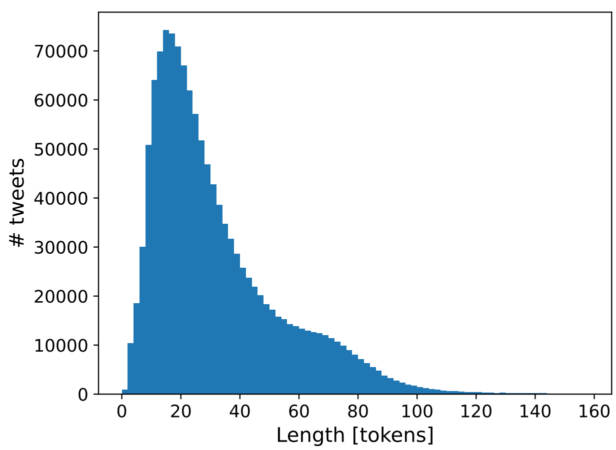
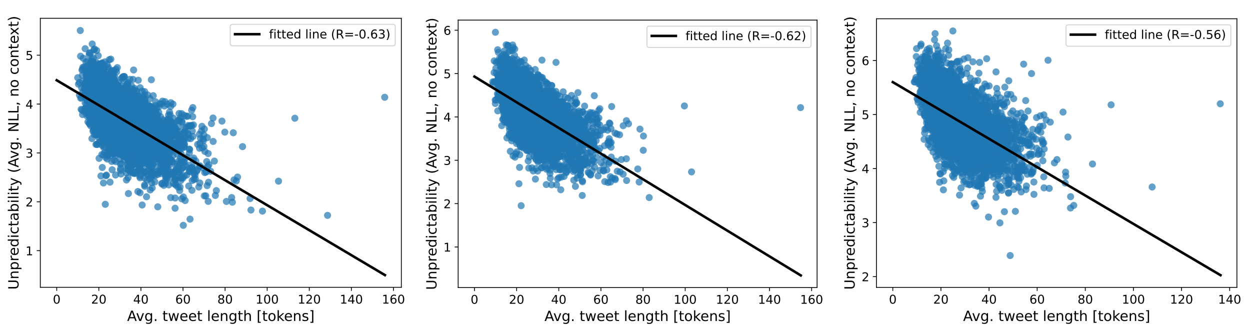


Figure 13: Distribution of tweet length in tokens (Llama-2 tokenizer). Tweets are from $\mathcal{T}^{\texttt{eval}}$

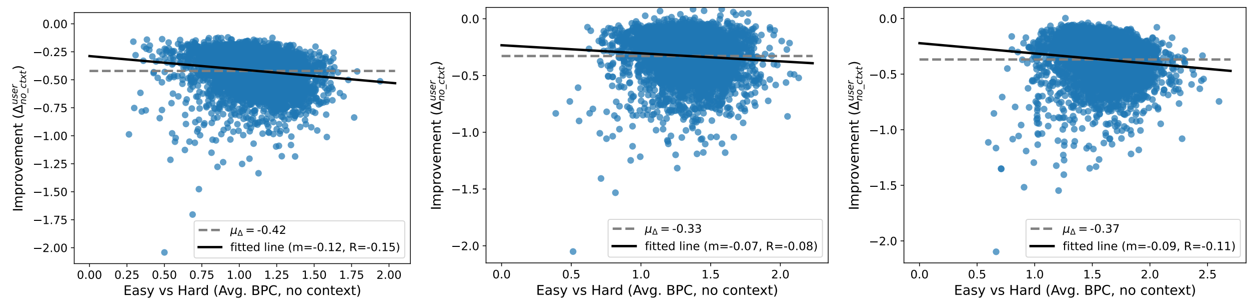


(a) Llama-2

(b) Falcon

(c) GPT-2-XL

Figure 14: Users with longer tweets are more predictable. On the y-axis, we plot the average NLL required to predict a user's tweets (no context setting). The x axis shows the user's average tweet length (in tokens).



(a) Llama-2 (b) Falcon (c) GPT-2-XL

Figure 15: Hard to predict users get more predictable with additional context, but only slightly. On the x-axis, we plot the average BPC required to predict a user's tweets (no context setting), effectively sorting them based on how predictable they are (easy vs. hard to predict). The y axis plots the relative improvement (decrease in BPC) with additional user context.

B.4 Two control groups: social and temporal control

Bagrow et al. (2019) introduce two control groups in their experiment: a *social* and a *temporal* control. Social control includes tweets from 15 random users. Temporal control on the other hand, selects 250 tweets that were authored around the same time as the tweets from the peer context. See Figure 10 for an illustration of both. The rugplot on the top of Figure 11 shows the same, but on real data of a random subject. Figure 16 shows our main results including both controls, while Figure 17 shows the corresponding effect sizes.

B.5 "Surprise effect" of first token

LLMs are known to have high negative log-likelihood on the first token of the input sequence. This could artificially inflate model uncertainty in the no context setting. To test its effect on our results, we omit predictions on the first token from our analysis. Figure 18 illustrates the results of this experiment. Indeed, the gap between the no context and other context settings shrink as expected. However the differences still remain significant and thus don't change our main findings.

B.6 Most improved token due to random context: @-mention

We were interested in which tokens benefited most after including random (i.e. social) context. We selected tokens with >100 occurences, and ranked them based on how much their prediction improved on average. Table 1 shows that the '@' token benefits the most across all three models, indicating that @-mentions are locations of greatest improvement. See Table 2 for a full table on top-10 most improved token predictions (with all possible contexts).

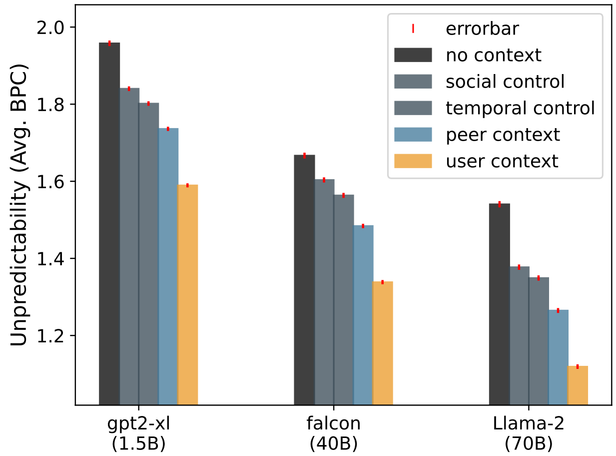
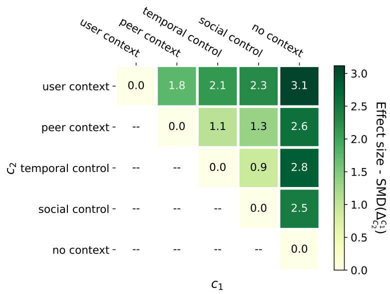
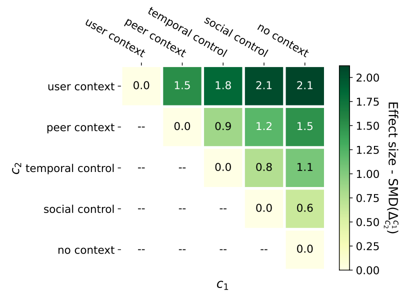


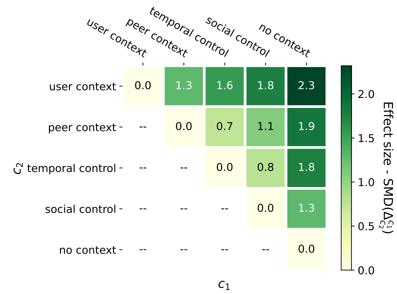
Figure 16: Average bits per character (BPC) required to predict user tweets, with 95% confidence intervals. Here both control settings are included: (i) past tweets from random users (social control, left) and (ii) past tweets made around the same time as the peer tweets (temporal control, right). These results are consistent with the ones presented in Fig. 1.



(a) Llama-2



(b) Falcon



(c) GPT-2-XL

Figure 17: Average effect size of c_2 relative to c_1 on user predictability. We look at the difference in predictability for each user $\Delta_{c2}^{c1}(u) = \bar{L}_u^{c1} - \bar{L}_u^{c2}$, where $\bar{L}_u^{c}$ is the average negative log-likelihood of user u under context c. Plotted values are standardized mean differences (SMD) of Δ_{c2}^{c1}. Darker green means greater improvement.

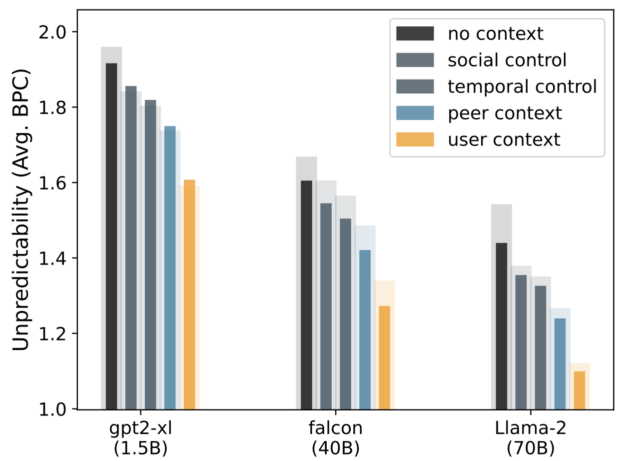
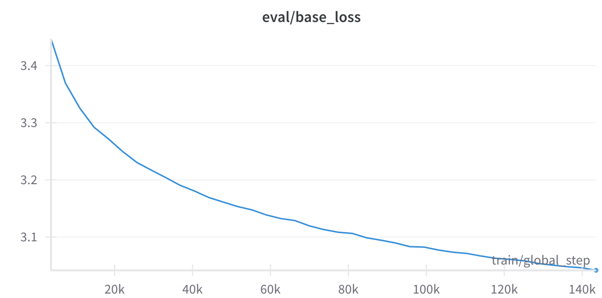


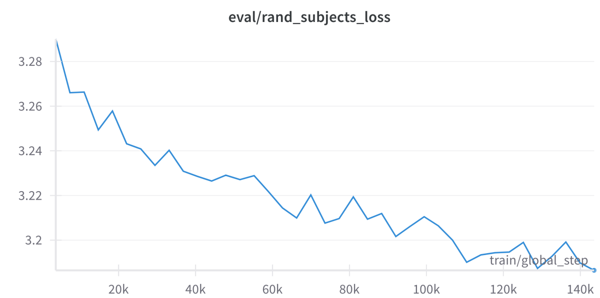
Figure 18: The effect of taking the first token out of the analysis. Overall, the average BPC decreases (predictability goes up), however the most significant drop is in the no context settings for GPT-2-XL and Llama-2. This happens because frequently the NLL of the first token (with no token preceding it) is usually quite high. Lighter bars are results from our original experiment.

	GPT-2-XL	Falcon	Llama-2
		social control	
0	@	@	▁@
1	âĢ	Brit	▁Hey
2	Ļ	ĠAmen	▁Ain
3	ĠARTICLE	ĠNah	▁Happy
4	ĠðŁij	ĠBru	▁Wait
5	ĠðŁ	ĠSame	▁tf
6	ĠâĢ	ĠDamn	/@
7	Ġâĺ	ĠWait	rach
8	Ġâľ	ĠDang	▁Okay
9	ľ	ĠNope	▁Ton

Table 1: Top ten tokens whose predictability went up the most (on average) after including tweets from random users as context (compared to no context). Notice how the @ sign is the token that got "bumped" the most, suggesting that the additional random context helped with predicting @-mentions. (Ġ is a special symbol for the space character in the GPT-2-XL and Falcon tokenizers.)



(a) Loss calculated on the 5% validation split ($\sim$ 0.5M tweets).



(b) Loss calculated on the evaluation set of 100 subjects (25k tweets). There is a lot more stochasticity, which in part is due to the smaller sample size and a less diverse pool of authors.

Figure 19: Loss curves of the pre-finetuning of GPT-2-XL on 10M tweets for 1 epoch. Validation loss (negative log-likelihood) was calculated on a 5% split. To make sure that pre-finetuning also improved subject predictability, we also selected a random subset ($n = 100$) of the 5k subjects. We used their evaluation tweets to calculate the loss and since each $|\mathcal{T}_u^{\texttt{eval}}| = 250$, this means a total of 25000 tweets.

C Finetuning

C.1 Pre-finetuning

With a batch size of 8, it took 147279 global steps to go over the entire set of training tweets once. In addition to the evaluation loss (NLL on the 5% validation split, which was checked periodically) we also tracked the loss calculated on the combined evaluation set of 100 subjects (25000 tweets in total) to make sure that the pre-finetuning improved prediction on our subjects as well. We present a figure of the loss curves in Figure 19.

	GPT-2-XL				Falcon				Llama-2			
	social	temporal	peer	user	social	temporal	peer	user	social	temporal	peer	user
0	@	@	hetti	ĠKraft	@	@	hyde	749	▁@	▁@	Extra	▁Kraft
1	âĢ	âĢ	DOM	hetti	Brit	Brit	perm	Ġgenealogy	▁Hey	▁Ain	DOM	medium
2	Ļ	Ļ	204	ĠARTICLE	ĠAmen	ĠNah	Agg	Ich	▁Ain	▁Hey	member	Extra
3	ĠARTICLE	ĠARTICLE	Agg	hyde	ĠNah	ĠBru	018	ildo	▁Happy	▁tf	zent	zent
4	ĠðŁij	ĠðŁij	rium	DOM	ĠBru	ĠAmen	641	hyde	▁Wait	Fil	*v*	*v*
5	ĠðŁ	ĠâĢ	alys	ĠSag	ĠSame	ĠDamn	atts	MCA	▁tf	▁Wait	Los	▁Via
6	ĠâĢ	ĠðŁ	perm	medium	ĠDamn	ĠWait	576	perm	/@	soft	members	hour
7	Ġâĺ	ĠPis	Hour	ĠRescue	ĠWait	ĠSame	931	ĠKeller	rach	rach	▁@	member
8	Ġâľ	ľ	hyde	ĠPis	ĠDang	ĠDude	Chi	ENV	▁Okay	▁Via	gat	DOM
9	ľ	Ġâĺ	Extra	ĠKeller	ĠNope	ellan	454	ĠVenue	▁Ton	▁Om	▁txt	ihe

Table 2: Top ten tokens whose predictability went up the most (on average) after including tweets from some context (compared to no context). Tokens with >100 occurences were selected.