Addressing the Ecological Fallacy in Larger LMs with Human Context

Nikita Soni[1], **Dhruv Vijay Kunjadiya**[1], **Pratham Piyush Shah**[1], **Dikshya Mohanty**[1], **H. Andrew Schwartz**[1,2], **and Niranjan Balasubramanian**[1]
[1]Department of Computer Science, Stony Brook University
[2]College of Connected Computing, Vanderbilt University, USA
{nisoni, niranjan}@cs.stonybrook.edu

Abstract

Language model training and inference ignore a fundamental linguistic fact: there is a dependence between multiple sequences of text written by the same person. Prior work has shown that addressing this form of *ecological fallacy* can greatly improve the performance of multiple smaller (~124M) GPT-based models. In this work, we ask if addressing the ecological fallacy by modeling the author's language context with a specific LM task (called HuLM) can provide similar benefits for a larger-scale model, an 8B Llama model. To this end, we explore variants that process an author's language in the context of their other temporally ordered texts. We study the effect of pre-training with this author context using the HuLM objective, as well as using it during fine-tuning with author context (*HuFT:Human-aware Fine-Tuning*). Empirical comparisons show that addressing the ecological fallacy during fine-tuning alone using QLoRA improves the performance of the larger 8B model over standard fine-tuning. Additionally, QLoRA-based continued HuLM pre-training results in a human-aware model generalizable for improved performance over eight downstream tasks with linear task classifier training alone. These results indicate the utility and importance of modeling language in the context of its original generators, the authors.

1 Introduction

To date, if one asks an LLM to complete the phrase, "Language is generated by __", they will get 'humans' or 'people' as the two most likely words to follow. Yet, the standard language modeling task itself does not model the dependence between the token sequences and the people behind language, an *ecological fallacy* of assuming text sequences from the same person are independent (or treated the same as those from different people). Language representations devoid of their originating human contexts lack the richness and variance that

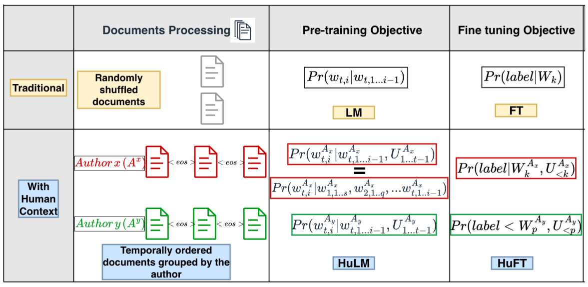

Figure 1: Addressing the *ecological fallacy* in larger LMs with human context involving processing documents, pre-training, and fine-tuning within the author's context. Human Language Modeling (HuLM) and Human-aware Fine-Tuning (HuFT) consist of training objectives conditioned on author's historical language as the human context (U).

natural human contexts bring (e.g. generated language lacks variance in expressed psychological traits (Giorgi et al., 2023; Varadarajan et al., 2025)), limiting model ability to address biases (Soni et al., 2024b).

In this work, we ask: *does addressing this ecological fallacy help large language models?* In particular, we explore the impact of processing language within the human context as modeled by the author's previous texts. Prior work showed that this ecological fallacy can be remedied through a *Human Language Modeling* (HuLM) task (Soni et al., 2022), which models the human behind the language by conditioning next-word prediction not only on the immediate discourse context but also on the broader human context via temporally ordered texts from the same author. Soni et al. show continued pretraining of a small scale GPT-2 variant (124M parameters) on this HuLM task improves performance both in terms of LM perplexity and downstream applications.

However, it is not clear *a priori* that the more powerful larger models need this additional human context. One may posit that LLMs with billions

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 121–144
July 3-4, 2026

of parameters trained over trillions of tokens, already capture language from a large population of humans and thus overcome any representational or distributional shortcomings that arise from the lack of processing in the author's context.

To address this, we investigate three different ways of incorporating human context into larger LMs, in terms of author's historical language: (i) directly including human context to the text being processed for downstream tasks and training a task-specific linear classifier, (ii) fine-tuning model parameters using QLoRA for downstream tasks by including human context to the text, which we call *HuFT: Human-aware Fine-Tuning*, and (iii) continuing pre-training model parameters using QLoRA by including human context (i.e., HuLM). We select open source Llama 3.1 8B model weights for our study, curate a new Large Human Language Corpus (LHLC) for continued HuLM pre-training, curate six cleaned and processed downstream task datasets with anonymized author identifiers, and evaluate the three ways to include human context over a total of eight downstream tasks. We scope our study to one model family as it is a resource intensive endeavor, requiring substantial compute and time.

Our empirical results highlight several important findings: (i) Human-aware QLoRA-based Fine-Tuning (HuFT) effectively includes human context in task specific settings resulting in improved performance, (ii) Continued HuLM pre-training (QLoRA-based) results in a human-aware model that can better generalize over multiple tasks with linear classifier training alone, however, (iii) training a task-specific linear classifier by directly including the human context proves to be ineffective for non human-aware models.

In summary, we make the following contributions in this work: 1) an empirical demonstration of the value in addressing the ecological fallacy in larger language models evaluated over eight downstream tasks, 2) trained a bigger HuLM model (in the range of 8B parameters) using QLoRA in multiple settings, and 3) developed a diverse and substantial HuLM data corpus consisting of texts from Reddit (Giorgi et al., 2024a; Liu et al., 2024), Blog Authorship Corpus (Schler et al., 2006), Twitter (Giorgi et al., 2024b; Soni et al., 2022), Gutenberg Books (Bejan, 2021), Amazon Product Reviews (Hou et al., 2024), and StackExchange (Lambert et al., 2023), as well as 4) curated six downstream tasks and datasets with author context via author's historical texts. We also present results with prompting, ablations and qualitative analysis to further support our findings. Datasets, models, and code is available on GitHub[1].

2 Related Work

A wealth of past work has shown the efficacy of looking at language within the larger context of who the author is (Soni et al., 2024a, 2025a) or their demographics in multiple applications, such as sentiment analysis (Mireshghallah et al., 2021), reducing social biases (Garimella et al., 2022), or mental health assessments (Varadarajan et al., 2024; Soni et al., 2025b). In the realm of large LMs, prior work has shown benefits in considering a person's dynamic emotional states (Ganesan et al., 2022; Singh et al., 2025) to generate empathetic dialogs (Wang et al., 2022), or enhancing personalized responses (Tan et al., 2025) by injecting memory within model parameters using multiple LoRA modules, inspired by human memory mechanisms (Zhang et al., 2025).

Past works (Soni et al., 2022, 2024b) suggest including human context within the LM pre-training task of next word prediction to address the ecological fallacy of the LM task. Soni et al. introduce the task of **Human Language Modeling (HuLM)**, where they predict the next word given the previous words and an additional author's context in terms of the author's prior language, formulated as:

$$Pr(\mathbf{W}_t|\mathbf{U}_{t-1}) = \prod_{i=1}^{n} Pr(w_{t,i}|w_{t,1:i-1}, \mathbf{U}_{1:t-1}) \tag{1}$$

such that the task is to predict the next word ($w_{t,i}$) of a document W_t, given the document's previous words ($w_{t,1:i-1}$) and a dynamic human context $U_{1:t-1}$ which models all previous words in all prior documents written by the same person.

At the same time, larger LMs have demonstrated remarkable performance in many tasks (OpenAI, 2023; Hendrycks et al., 2021; Jimenez et al., 2024; Singhal et al., 2022). However, larger LMs have not yet been evaluated for the effects of processing language within a human context (i.e., an author's historical language) when continued to pre-train or when fine-tuned for downstream tasks (Matero et al., 2021). In this study, we assess the impact of addressing the ecological fallacy in larger LMs by

[1] https://github.com/soni-n/Larger_LMs_with_Human_Context

extending the concept of HuLM into larger LMs in terms of continued QLoRA pre-training and fine-tuning for downstream tasks.

3 Models and Methodology

We seek to evaluate the effectiveness of processing language within the author's context, i.e., mitigating the *ecological fallacy*, in larger LMs. We include this human context in three ways: i) training a task classifier using pre-trained embeddings, ii) fine-tuning model parameters for a downstream task, and iii) continue to pre-train for the human language modeling task.

We select Llama 3.1 8B (Touvron et al., 2023) as our base model. It adopts a decoder-only transformer architecture, allowing us to easily adapt it to the autoregressive HuLM task. Similar to Soni et al. (2022), we include the human context by collectively processing language written by the same author in the three ways we listed above. First, we discuss the difference in processing data traditionally and within the human context (i.e., within the author's context). Next, we describe the three ways that we use to include this human context and assess its impact on downstream tasks (refer Figure 1). Finally, we define the baseline models and details of the training setup.

Processing Data within the Human Context. Traditionally, documents are randomly shuffled and processed independently for the language modeling task. To mitigate the ecological fallacy in language processing, we process documents generated by a common source (i.e., the author) together, thus inducing dependence on this hierarchical source. In practice, we concatenate documents written by the same author, separated by the special *eos* token, ordering them temporally when the created time information is available. This approach of processing documents within the author's context remains consistent for both pre-training and fine-tuning with the human context.

For document-level labeled downstream tasks, where the task is to predict an outcome associated with a document, the models process all the tokens in the concatenated sequence and use the target document's last token's hidden states (of the last layer) to predict the associated outcome label, e.g., the author's stance or sentiment. For person-level labeled downstream tasks, where the task is to predict a personal attribute using a collection of documents written by the person, the models process the author's language similarly and use the averaged token embeddings across all tokens from an author to predict/estimate the associated label, e.g., the author's occupation or age.

Dataset	Epochs	Users	Docs *(millions)*	Tokens *(millions)*	UTF-8 bytes *(GB)*
Amazon	1	30,450	2.252	204.94	0.84
Blogs	3	21,180	0.303	85.81	0.34
Books	1	29,535	0.030	258.19	1.07
Twitter	3	20,110	2.290	62.17	0.23
Reddit	3	30,921	1.407	166.38	0.69
StackExchange	1	15,207	0.557	82.52	0.35
Total		**147,403**	**6.839**	**860.00**	**3.52**

Table 1: Subset of Large Human Language Corpus (LHLC) used as our pre-training data. The counts in this table are reflective of the numbers resulting after processing the data for multiple author-instances. Token counts are based on LLaMA 3.1 tokenizer.

Classifier Training with the Human Context. In this approach, we use a pre-trained model's embeddings, precisely the hidden states from the model's last layer, for documents processed within the author's context. These are processed in accordance to document- or person- level downstream tasks described above and used as inputs to a linear classifier which is trained for specific downstream tasks.

Human-aware Fine-Tuning: HuFT. In this method, we train the task-specific classifier as well as fine-tune the model parameters for specific downstream tasks. We use a similar approach to process the input based on the type of downstream task using the corresponding hidden states of the author's language derived from the model's last layer.

Continued HuLM Pre-training: HU-Llama. In this approach, we continue to pre-train Llama 3.1 8B for the next word prediction task in HuLM style using our curated dataset: LHLC (see Section 4.1). We train over concatenated sequences of temporally-ordered documents authored by the same person, instead of randomly sampling documents and processing them independently. We use the special token *eos* as the delimiter between concatenated documents. This results in each instance representing an author, inducing an explicit author's context by introducing dependence on the language written by the same author. This follows a flattened version of the data processing and HuLM pre-training task from Soni et al. (2022), resulting in HU-Llama (Human-aware Llama).

Task	Users	Docs	Labels	wpd$_{med}$	wpd$_{max}$	dpa$_{med}$	dpa$_{max}$	wpa$_{med}$	wpa$_{max}$
Occupation(blogs)	3,539	47,500	3,539	95	350	8	337	1166	4999
Age(blogs)	15,942	210,253	15,942	100	350	9	353	1189	4999
Movie Reviews(amazon)	16,369	109,213	109,213	119	350	6	17	826	3783
Business Reviews(yelp)	35,025	216,166	216,166	133	350	6	8	848	2090
Electronics Reviews(amazon)	15,026	130,316	130,316	67	350	9	12	719	2706
Book Reviews(amazon)	18,475	156,218	156,218	122	350	8	15	1077	3293
Sentiment(twitter)	10055	47280	10485	16	192	1	199	21	3478
Stance(twitter)	2349	17411	3021	15	33	1	2095	19	40018

Table 2: Curated two person-level (Occupation and Age) and four document-level task datasets. Statistics reported above, where wpd = words per document, dpa = documents per author, wpa = words per author, and med = median. Stance and Sentiment are two existing document-level datasets.

Baselines: Traditional Fine-tuning (TFT), Llama, Llama$_{\text{LHLC}}$. We use traditional fine-tuning (TFT) as the baseline to compare against HuFT. Here, we adopt the standard fine-tuning approach where the model is given an individual document and asked to predict the label. For document-level tasks, this translates to not using the author's context. For person-level tasks, we adopt a prior common approach (Soni et al., 2022) of fine-tuning the model by independently predicting a person-level attribute for each document written by the author and taking an average or the mode of the predictions across all documents to arrive at a person-level prediction.

We compare the bigger HuLM model: HU-Llama, with its counterpart non-HuLM model, Llama, across downstream tasks using classifier training and human-aware fine-tuning (HuFT). While HuLM models are naturally designed to adopt HuFT, smaller LMs are not known to use this approach, usually due to the limitation of context lengths. However, larger LMs that can process larger contexts have not been evaluated for HuFT. So in addition to the HuLM model (HU-Llama), we also evaluate Llama (non-HuLM model) using HuFT for downstream tasks, leveraging their capacity to process long contexts.

Further, to tease apart the impact of our pre-training data corpus—LHLC—, we continue to pre-train Llama 3.1 8B for the standard next word prediction task using our LHLC dataset and call it Llama$_{\text{LHLC}}$. We compare our methodologies of including human context with Llama$_{\text{LHLC}}$ as well.

Training Setup: QLoRA Training, and Experimental Settings. For both, continued pre-training and fine-tuning the models, we use low-rank adapters and 4-bit quantized weights (QLoRA) (Dettmers et al., 2023) accommodating for compute availability. Additionally, we use mixed-precision training, performing operations in half-precision format to speed up computation, as well as use Accelerate (Gugger et al., 2022) for a distributed environment in some time-consuming experiments. We use the PEFT library (Han et al., 2024) integrated in the Hugging Face (HF) library (Wolf et al., 2020) to train QLoRA weights associated with Q, K, V, O of the self-attention mechanism.

At the 8B scale, this setup enables us to continue HuLM pre-training with a batch size of 3 per GPU on our hardware (NVIDIA H100 80GB GPUs) using 8192 tokens per author-instance. Similarly, using ditto document-tokens, we use a batch size of 123 with each instance (representing each document) limited to 200 tokens for Llama$_{\text{LHLC}}$ continued pre-training, where each document is processed independently. We run initial experiments with smaller data samples using learning rates 1e-6, 3e-4, and 5e-5, and resort to the best performing, 3e-4, when training with full data. We train on full data incrementally, and Table 1 details the number of epochs for which each data source was trained. Finally, we merge the QLoRA pre-trained HuLM- and traditional-LM- adapters into the respective base Llama models, yielding HU-Llama and Llama$_{\text{LHLC}}$. We use a similar QLoRA setup for HuFT and TFT starting with the resulting pre-trained models. We train all models for 5 epochs or 10 epochs with a learning rate of 3e-4 and an early stopping threshold set to 6 on the evaluation loss. We cap the training tokens to 4096 per instance (i.e., per author) and optimize training using cross-entropy loss for classification tasks and mean squared error loss for regression tasks.

We run statistical significance testing using paired t-test for regression tasks and permutation test for classification tasks.

Task	Train			Dev			Test		
	Users	Docs	Labels	Users	Docs	Labels	Users	Docs	Labels
Occupation(blogs)	2,135	28,199	2,135	532	7,770	532	872	11,531	872
Age(blogs)	10,354	135,594	10,354	2,402	32,773	2,402	3,186	41,886	3,186
Movie Reviews(amazon)	10,621	71,338	71,338	2,460	16,168	16,168	3,288	21,707	21,707
Business Reviews(yelp)	22,856	141,153	141,153	5,233	32,179	32,179	6,936	42,834	42,834
Electronics Reviews(amazon)	9,768	84,972	84,972	2255	19,527	19,527	3,003	25,817	25,817
Book Reviews(amazon)	11,974	101,614	101,614	2,811	23,565	23,565	3,690	31,039	31,039
Sentiment(twitter)	6,246	28,808	6,461	1,000	4,548	1,030	2,859	13,924	2,994
Stance(twitter)	1,361	11,318	1,658	332	1,996	418	768	4,097	945

Table 3: Train, Dev, and Test split statistics for each dataset across tasks, including number of users, documents, and labels. Here, we use the splits from SemEval tasks for stance and sentiment (Nakov et al., 2013; Mohammad et al., 2016), and the author's context from Lynn et al. (2019); Soni et al. (2022). For the person-level tasks, we stratify on the number of words per user and maintain a consistent label-proportions and no overlapping authors in each split.

4 Datasets and Tasks

4.1 Pre-training Data: Large Human Language Corpus (LHLC)

Human language modeling requires pre-training datasets that can provide language in the author's context, i.e., where text can be attributed to its source (author) while maintaining the privacy of a person's identity. Despite the abundant availability of datasets with meta-data consisting of *anonymous user identifiers*, to the best of our knowledge, there is no cleaned and processed dataset available to use directly. To facilitate progress in human language modeling and personalized modeling research, we build and release the first version of our pre-training dataset, LHLC—a large, multi-source corpus containing millions of documents across more than 140K authors (resulting in more than 170K author instances when an author's language is used in more than one input instance), and planned release of a data report consisting the details of our dataset construction and design principles[2]. Briefly, the curation process for LHLC is inspired by Soldaini et al. (2024) and the steps include: 1) removing missing data, 2) data deduplication, 3) english filtering, 4) text formatting (encoding, URLs, etc.), 5) toxicity filtering, 6) anonymization, 7) creating multiple author-instances for certain authors. Here, we use a subset of this data summarized in Table 1, where we include counts based on multiple author-instances. We present the original counts for the authors and documents in the Appendix Table 10. We consider a mix of domains with a stronger focus on social language (e.g., blogs, Reddit, Twitter) along with more topic-focused domains (e.g., Amazon Products Reviews, StackExchange) and domains with less informal language (e.g., books) to regularize training. Details of the training split can be found in the Appendix Tables 11 and 12.

4.2 Downstream Datasets and Tasks

We consider eight downstream tasks that involve assessing person-level attributes grouped into two categories based on the prediction level: **document-level** and **person-level**. In the first category, given a target text sequence (document) written by a person, the model predicts a label associated with that document (e.g., the person's stance on a topic such as *atheism*). In the second, given multiple text sequences (documents) written by the same person, the model predicts/estimates a label representing a broader personal attribute (e.g., occupation or age). Train, dev, and test splits for all tasks can be found in Table 3

We select existing datasets as well as curate new datasets for downstream tasks, similar to LHLC curation, while also stratifying labels distributions (refer to Table 2 for downstream data statistics and Appendix Table 13 for downstream task labels proportions). The tasks considered assist in evaluating the impact of including the human context while processing language. Although it may not seem obvious how human context can be effective for document classification tasks since traditionally such tasks have relied on linguistic cues within a single document. However, such tasks can leverage the richness and variance provided by the author's historical context, for instance disambiguating sarcasm, idiosyncratic metaphors, or personal style, which we discuss further as part of qualitative analysis in Section 5.4. Furthermore, all selected tasks

[2]Report release to be announced on project GitHub page: https://github.com/soni-n/Larger_LMs_with_Human_Context

Model	Include HC	Document-Level Cls ($F1$) Movie	Business	Book	Elec	Stn	Sent	Person-Level Cls ($F1$) Occ	Reg (r) Age
Llama	No	57.99	65.99	59.90	59.06	64.29	62.01	47.48	0.826
Llama	Yes	55.06	63.37	57.80	57.61	62.96	61.40	53.59	0.853
HU-Llama	**Yes**	**<u>58.48</u>**†	**<u>66.80</u>***†	**<u>61.88</u>***†	**<u>60.93</u>***†	**<u>67.08</u>**†	**<u>62.93</u>**†	**<u>54.73</u>***	**<u>0.858</u>***†

Table 4: Evaluating directly including human context (HC) using pre-trained embeddings and training task-specific linear classifiers. Results are reported in weighted F1 for classification (cls) tasks and in pearson r for regression (reg) task. Bold indicates best in column, * and † indicate statistical significance $p < .05$ w.r.t not including human context, and w.r.t including human context using Llama embeddings respectively . Underline indicates where continued QLoRA HuLM pre-training additionally helps.

Model	FT style	Document-Level Labels Cls ($F1$) Movie	Business	Book	Elec	Stn	Sent	Person-Level Labels Cls ($F1$) Occ	Reg (r) Age
Llama	TFT	66.40	70.93	65.32	65.26	**71.37**	**76.44**	52.45	0.882
Llama	**HuFT**	**67.52***	**74.36***†	69.22	**68.84***†	69.09	76.07	52.28	**0.916***
HU-Llama	**HuFT**	67.05	73.36	**<u>69.43</u>***	68.22	70.77	75.25	**<u>57.50</u>***†	**0.916***

Table 5: Evaluating QLoRA based Human-aware Fine-Tuning (HuFT). Results reported in weighted F1 for classification (cls) and in pearson r for regression (reg) tasks. Bold indicates best in column, * and † indicate statistical significance $p < .05$ w.r.t traditional fine-tuning (TFT), and w.r.t second best HuFT results respectively. Underline indicates where continued QLoRA HuLM pre-training additionally helps.

have real-world applications such as sentiment analysis, bias mitigation, enforcing safety policies (e.g., flagging underage users in sensitive domains), tailoring educational content, analyzing labor market trends, and authorship attribution. At the same time, we acknowledge the potential dual-use risks of such modeling and person associated tasks—such as profiling, manipulation, stereotyping, or targeted advertising—and emphasize the need for data consent and transparency. We discuss these considerations in greater detail in Section 6.

Person-Level. Similar to LHLC corpus, we curate two person-level downstream task datasets and training splits from existing blogs (Schler et al., 2006) corpus. One requires classifying a person's **occupation**, and the other requires estimating a person's **age**, given the blogs written by them for both tasks. We limit the occupation classification data to consist of the top 5 occupations from the blogs corpus (student, technology, arts, communication and media, and education).

Document-Level. We curate four document-level downstream task datasets from existing sources—Amazon movie Reviews (McAuley and Leskovec, 2013), Yelp Business Reviews (Asghar, 2016), Amazon Electronics Reviews (Hou et al., 2024), and Amazon Books Reviews (Hou et al., 2024). All four tasks involve predicting a rating (ranging 1 to 5) that a person would give corresponding to a review written by them. We select different domains and products to diversify the task and assess performance impact across domains and topics. Further, we select two publicly available datasets consisting of authors' context (Soni et al., 2022) for the **stance detection** and **sentiment analysis** tasks from SemEval (Nakov et al., 2013; Mohammad et al., 2016) using ditto train, validation, and test splits. There are five sub-datasets corresponding to five topics (atheism, abortion, climate, feminism, and Hillary Clinton) for the stance detection task, which requires predicting the stance (for/against/neutral) of a person for a particular topic given a tweet written by them. Similarly, for sentiment analysis, the task is to predict the sentiment (positive/negative/neutral) of a person given a tweet written by them. For both tasks, the author's context, where available, consists of historically written (unlabeled) tweets by the author.

5 Results and Discussion

In this section, first we describe the main results of studying the impact of incorporating human context in larger LMs in three settings: (i) classifier-only training, (ii) QLoRA-based Human-aware Fine-Tuning (HuFT), and (iii) QLoRA-based con-

Model	FT style	Document-Level Labels Cls ($F1$)						Person-Level Labels Cls ($F1$)	Reg (r)
		Movie	Business	Book	Elec	Stn	Sent	Occ	Age
Llama	TFT	57.99	65.99	59.90	59.06	64.29	62.01	47.48	0.826
LlamaLHLC	TFT	58.31	65.64^{o}	61.18^{o}	60.22^{o}	**69.25**^{o}	**63.17**	50.61^{o}	0.835^{o}
Llama	**HuFT**	55.06	63.37	57.80	57.61	62.96	61.40	53.59	0.853
LlamaLHLC	**HuFT**	55.43	63.57	59.36$^{\hat{}}$	57.54	66.37$^{\hat{}}$	62.38	54.01	0.855$^{\hat{}}$
HU-Llama	**HuFT**	**58.48**†	**66.80***†	**61.88***†	**60.93***†	67.08	62.93	**54.73***	**0.858***

Table 6: Evaluating the effect of our LHLC data corpus on directly including human context using task classifier-only training. Results reported in weighted F1 and pearson r. Bold indicates best in column. We indicate statistically significant ($p < .05$) difference in results between: LlamaLHLC HuFT and HU-Llama HuFT with †, LlamaLHLC TFT and HU-Llama HuFT using *, LlamaLHLC TFT and Llama TFT with ^{0}, and LlamaLHLC HuFT and Llama HUFT using $\hat{}$.

Model	Include HC	Document-Level Labels Cls ($F1$)						Person-Level Labels Cls ($F1$)	Reg (r)
		Movie	Business	Book	Elec	Stn	Sent	Occ	Age
Llama-Instruct	No	**61.14***	59.17	**60.16***	59.72	68.44	**65.75***	-	-
Llama-Instruct	Yes	59.70	**60.34***	58.63	**61.16***	**70.35***	64.81	37.09	0.145

Table 7: Evaluating directly including human context (HC) in prompting using Llama-Instruct 3.1 8B model. Results reported in weighted F1 for classification (cls) and in pearson r for regression (reg) tasks. Bold indicates best in column, * indicates statistical significance $p < .05$ between not including human context and including human context for task-specific prompting.

tinued HuLM pre-training, presented in Tables 4 and 5. Next, we summarize additional results supporting our main investigative study with ablations, interpretive experimental and qualitative analyses. Finally, in section 5.5 we discuss the inferences and implications for the observations stated.

5.1 Including human context with task-specific classifier-only training

We find that directly including human context using Llama embeddings, for task-specific classifier-only training, improves performance on person-level tasks (estimating age and classifying occupation). However, it does not benefit any of the document-level tasks (see Table 4, rows "Llama with human context" and "Llama without human context"). At the same time, directly including human context using HU-Llama embeddings—which is Llama continued to pre-train with human context for HuLM task using QLoRA—for task-specific classifier training yields substantial gains across all the document- and person-level tasks. Five out of eight tasks, including both person-level tasks, show statistically significant results when including human context as compared to not including human context.

5.2 Including human context in downstream task fine-tuning: HuFT

We find that human-aware task-specific fine-tuning Llama using QLoRA (see HuFT in Table 5) shows substantial performance gains over fine-tuning without human context (TFT) in six out of eight downstream tasks with statistical significance ($p < .05$). Two document-level tasks—stance and sentiment—does not benefit in HuFT setting, however, have no statistically significant difference from TFT setting. We also note here that these two tasks have only a few instances with historical author context.

5.3 Including human context in continued pre-training: HuLM

While continued QLoRA HuLM pre-training in task-specific classifier-only training setting shows benefits (see HU-Llama in Table 4), we also see that these gains are surpassed by Llama with human-aware QLoRA task fine-tuning in three tasks (Movie, Business, and Electronics Reviews) with two having statistically significant results (see Table 5). Two other tasks (occupation, Book Reviews Rating) sustain best performances with HU-Llama in HuFT settings and one other (estimating age) is at par with Llama HuFT.

Model	FT style	Document-Level Cls ($F1$) Business(F1) RS42	RS3	RS1234	Sentiment(F1) RS42	RS3	RS1234	Person-Level Cls ($F1$) Occ(F1) RS42	RS3	RS1234	Reg (r) Age(F1) RS42	RS3	RS1234
Llama	TFT	70.93	71.70	71.24	**76.44**	76.35	**79.05***	52.45	50.71	50.21	0.882	0.881	0.885
Llama	HuFT	**74.36***†	**74.12***†	73.71	76.07	76.86	76.44	52.28	55.65	55.34	**0.916***	**0.913***	**0.914***
HU-Llama	HuFT	73.36	73.48	**74.19***†	75.25	**78.05***	75.32	**57.50***†	**56.76***	**55.82***	**0.916***	**0.913***	**0.914***

Table 8: QLoRA-based HuFT and TFT experiments with different random seeds (RS) for selected downstream tasks. Results in weighted F1 for classification (cls) tasks and in pearson r for regression (reg) task. Bold indicates best in column, * and † indicate statistical significance $p < .05$ w.r.t TFT and w.r.t second best HuFT results respectively.

5.4 Ablation and Interpretive Analyses

Effect of LHLC dataset. Training on a large dataset, such as LHLC, could itself explain part of the performance gains we observe on downstream tasks with HU-Llama in classifier-only training setting. Table 6 shows continued pre-training (QLoRA-based) on LHLC for the standard LM task (Llama$_{\text{LHLC}}$) fares worse compared to HU-Llama on all eight downstream tasks in the HuFT setting, and on six tasks in the TFT setting. Conversely, Llama$_{\text{LHLC}}$ shows better or at par performance as compared to Llama in both TFT and HuFT settings showing the benefits of continued LM pre-training on LHLC.

We perform a similar analysis to evaluate the effect of LHLC on QLoRA-based HuFT setting and report results in the Appendix Table 14 with similar findings except minor mixed results for Llama$_{\text{LHLC}}$ verus Llama, albeit, following performance trends as seen in Table 5.

Including human context in prompting. We further evaluate directly including human context in prompting using Llama-Instruct 3.1 8B model (prompts and details in Appendix A.4). We find mixed results in directly including author's historical language with zero-shot prompting (see Table 7[3]. We also experiment with limited historical context and historical context along with labels over two selected downstream tasks but notice only marginal differences (see Appendix Table 17). Overall, Llama-Instruct 3.1 8B model is not able to effectively use the human context with prompting.

Experiments with randomness. We experiment with two other random seeds (3 and 1234) to observe the effect of randomness on selected downstream tasks. We find similar performance results as our original random seed (42) on selected tasks with an exception where HuFT performs better than TFT for sentiment analysis with one random seed (see Table 8). These results are consistent with the main findings of HuFT improving downstream task performance regardless of the randomness.

Qualitative Analysis We follow a structured approach by looking at error in predictions (detailed process listed in Appendix A.5) to manually review examples where human context as the author's historical language helps or hurts a model's predictions. We find examples where models in TFT setting make the wrong prediction because the target text does not directly reveal the stance or occupation of a person whereas models in HuFT setting benefit from the historical human context to predict correctly. Conversely, we also find cases where the historical language can be misleading causing the models to mispredict in the HuFT setting. We show a few examples in Table 9 and provide task-wise qualitative analysis in the Appendix A.5.

5.5 Discussion

Human-aware QLoRA task fine-tuning proved effective for improving performance by including human context for task-specialized models where model parameter tuning is feasible. Conversely, including human context in continued pretraining, even within the QLoRA environment, yielded a human-aware model that generalized better across multiple tasks with linear classifier training alone. At the same time, directly including human context in the classifier-only training setting was ineffective for non-human-aware models, consistent with the relatively poorer results we observe when prompting with human context. We hypothesize that the reason could be the inability of Llama 3.1 8B model to effectively use a large amount of human context to improve its performance on downstream tasks. This is further supported by the better results of Llama in TFT setting for the stance and sentiment analysis tasks, which are the only two tasks where sufficient author historical context is not available. Additionally, as our qualitative analysis showed, au-

[3]Since it is not natural to prompt for person-level tasks with individual texts (i.e., no human context), we skip those in the results table.

Task	Target Text	Selected Human Context	Why Human Context Helps/Hurts
Stance (Atheism) True: FOR HuFT: FOR TFT: AGAINST	I #DenounceHarper for refusing to include family planning in foreign aid even though spending $1 could save $6 #wherestheFP	[...] bringing in religious groups [...] to access our kids in secrecy. [. . .] anti-choice [..] denying accurate SexED [..] harm they cause. [. . .]	The user's history repeatedly links policy to religious influence and Sex ED, clarifying the stance.
Occupation True: Tech. HuFT: Tech. TFT: Arts	". . . homeward bound :: need my hermit shell . . . " **[TFT → Technology]** [...] "[. . .] but the greatest [. . .] CPU [. . .] is the human brain [. . .] how you loop them [. . .]" **[TFT → Arts]**	i am mack, i will some day become great [. . .] unused CPU power [. . .] banning evolution from science classes [. . .] reactions to NASA discoveries and scientific progress [. . .]	TFT can be right sometimes but majorly predicts "arts". Repeated references to computing metaphors, science education, and technology–policy reasoning help HuFT.
Business review True: 4.0 HuFT: 1.0 TFT: 4.0	I've never had Chinese food before, and after the first bite, "wow this tastes like it was just cooked for me." This happened every time [...]	I would never eat here. (0.0) [. . .] no thanks. (1.0) [. . .] a lot really pissed me off. (2.0)	History is dominated by long, critical complaints , leading the model to discount concise praise and predict too low.

Table 9: Selected examples for qualitative analysis that demonstrate how including author's historical language as human context can improve model predictions, and in some cases, mislead them, along with plausible reasons.

thors' historical context can be ambiguous in some instances. Such results suggest a future research study on retrieving relevant historical language to provide an effective human context.

These findings should be taken in light of several limitations: the acutely low number of model parameters (~0.17%) that could be trained within the QLoRA setup, as well as the smaller size of our LHLC dataset relative to that of the original Llama pre-training data. In addition, our study comprehensively experiments with a specific family and size of models, in part because these experiments are extremely resource intensive, requiring substantial compute and time. This calls upon future studies to experiment with full or higher percentages of model parameter tuning, and to assess different model families and sizes. Further, we highlight the importance of smaller sized (~1B, ~8B) models becoming human-aware, as many use cases involve sensitive data and smaller models can be hosted locally, thereby preserving user consent and privacy.

6 Conclusion

Scaling has delivered impressive advances for language models. However, these models ignore the larger dependence between sequences of text that come from the same person. This work studied the impact of remedying this issue in large-scale language models (with 8B parameters) by modeling the author's prior language contexts. A simple change to the target task fine-tuning, where we incorporate the author's prior language, led to significant improvements over standard ways of fine-tuning for task-specialized models. Pretraining with author context based language modeling (HuLM) on our curated Large Human Language Corpus (LHLC) yields a human-aware model (8B parameters) that can provide generalized benefits over multiple tasks with simply training a linear task classifier. These results together demonstrate the utility of modeling the primary generators of language, humans, in large language models.

Limitations

The purpose of our study is to consider the effects of processing language within the author's context in larger LLMs within the scope of continued pre-training and fine-tuning. We resort to quantized low rank adaptation of some model parameters as we are limited by the compute availability. This may result in reduced efficacy of the continued pre-training of the HuLM task within larger LMs. Thus, we note that assessing the full impact of HuLM pre-training in larger LMs remains an open question. Additionally, we note that the author's context may be dependent on the quality of the text documents used from their previously written language. This is a yet another research question remaining to be explored and beyond the scope of our study. Furthermore, our study's scope does not include prompt engineering or comprehensive assessment of the efficacy of different prompting strategies in various conditions with the author's context. We include basic prompting experiments with selected experiments on limited historical context and labeled historical context for the sake of completeness of comparison only.

Ethical Considerations

The multi-level human-document-word architecture of HuLM enables large language models to incorporate dependencies across an individual user's prior language, rather than treating each text sample in isolation. This shift toward modeling the human generators of language unlocks new potential for improving fairness, personalization, and contextual understanding. However, the same capability that allows for richer user-level context also raises important ethical concerns—particularly regarding the risks of misuse, such as behavioral profiling or manipulation based on language history.

To mitigate these risks, we systematically review each dataset incorporated into the corpus, identifying and removing user identifiers. This process was followed by thorough manual checks to ensure that no personally identifiable information remained. These safeguards were essential for protecting user privacy and reducing the likelihood of unintended exposure of sensitive information from social media content.

Additionally, our models/architecture doesn't explicitly rely on or encode user attributes during pretraining. By focusing solely on patterns in language use—rather than incorporating static user-level features—we aim to preserve privacy while still capturing the richness of human communication. This approach aligns with our broader objective of building ethically responsible, human-centered language models.

Acknowledgments

This work was supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005, a grant from the CDC/NIOSH (U01 OH012476), and a grant from the NIH/NIMH (CREATE: Center for Advancing Therapy with AI; P50 MH 139450). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, any other government organization, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *CoRR*, abs/1605.05362.

Matei Bejan. 2021. 15000 gutenberg books. https://www.kaggle.com/datasets/mateibejan/15000-gutenberg-books.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Adithya V Ganesan, Vasudha Varadarajan, Juhi Mittal, Shashanka Subrahmanya, Matthew Matero, Nikita Soni, Sharath Chandra Guntuku, Johannes Eichstaedt, and H Andrew Schwartz. 2022. Wwbp-sqt-lite: Multi-level models and difference embeddings for moments of change identification in mental health forums. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 251–258.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 311–319, Online only. Association for Computational Linguistics.

Salvatore Giorgi, Kelsey Isman, Tingting Liu, Zachary Fried, João Sedoc, and Brenda Curtis. 2024a. Evaluating generative ai responses to real-world drug-related questions. *Psychiatry Research*, 339:116058.

Salvatore Giorgi, Kelsey Isman, Tingting Liu, Zachary Fried, João Sedoc, and Brenda Curtis. 2024b. Evaluating generative ai responses to real-world drug-related questions. *Psychiatry Research*, 339:116058.

Salvatore Giorgi, David M. Markowitz, Nikita Soni, Vasudha Varadarajan, Siddharth Mangalik, and H. A. Schwartz. 2023. "i slept like a baby": Using human traits to characterize deceptive chatgpt and human text. In *IACT@SIGIR*.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik

Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? *Preprint*, arXiv:2310.06770.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. 2023. HuggingFace H4 Stack Exchange Preference Dataset. https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024. Aligning large language models with human preferences through representation engineering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10619–10638, Bangkok, Thailand. Association for Computational Linguistics.

Veronica Lynn, Salvatore Giorgi, Niranjan Balasubramanian, and H Andrew Schwartz. 2019. Tweet classification without the tweet: An empirical examination of user versus document attributes. In *Proceedings of the third workshop on natural language processing and computational social science*, pages 18–28.

Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H Andrew Schwartz. 2021. Melt: Message-level transformer with masked document representations as pre-training for stance detection. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 2959–2966.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908.

Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim, and Dimitrios Dimitriadis. 2021. Useridentifier: Implicit user representations for simple and effective personalized sentiment analysis. *CoRR*, abs/2110.00135.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. https://openai.com/research/gpt-4.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Khushboo Singh, Vasudha Varadarajan, Adithya V Ganesan, August Håkan Nilsson, Nikita Soni, Syeda Mahwish, Pranav Chitale, Ryan L Boyd, Lyle Ungar, Richard N Rosenthal, et al. 2025. Systematic evaluation of auto-encoding and large language model representations for capturing author states and traits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18955–18973.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Nikita Soni, Niranjan Balasubramanian, H Andrew Schwartz, and Dirk Hovy. 2024a. Comparing pre-trained human language models: Is it better with human context as groups, individual traits, or both? In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 316–328.

Nikita Soni, Pranav Chitale, Khushboo Singh, Niranjan Balasubramanian, and H. Andrew Schwartz. 2025a.

Evaluation of LLMs-based hidden states as author representations for psychological human-centered NLP tasks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7673–7682, Albuquerque, New Mexico. Association for Computational Linguistics.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Nikita Soni, August Håkan Nilsson, Syeda Mahwish, Vasudha Varadarajan, H Andrew Schwartz, and Ryan L Boyd. 2025b. Who we are, where we are: Mental health at the intersection of person, situation, and large language models. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 300–313.

Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024b. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8646, Mexico City, Mexico. Association for Computational Linguistics.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2025. Democratizing large language models via personalized parameter-efficient fine-tuning. *Preprint*, arXiv:2402.04401.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Vasudha Varadarajan, Salvatore Giorgi, Siddharth Mangalik, Nikita Soni, Dave M Markowitz, and H Andrew Schwartz. 2025. The consistent lack of variance of psychological factors expressed by llms and spambots. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pages 111–119.

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, et al. 2024. Archetypes and entropy: Theory-driven extraction of evidence for suicide risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 278–291.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4634–4645, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kai Zhang, Yejin Kim, and Xiaozhong Liu. 2025. Personalized llm response generation with parameterized memory injection. *Preprint*, arXiv:2404.03565.

A Appendix

A.1 Training Dataset Additional Details

Table 10 presents the original number of users and documents in our pretraining dataset, as curated before going through the processing of creating multiple author-instances. Table 11 shows the train, dev, and test splits for our pretraining data reflecting the counts resulting post processing for multiple author-instances, while Table 12 present the same details with the original number of authors and documents.

Dataset	**Epochs**	**Users**	**Docs** *(millions)*	**Tokens** *(millions)*	**UTF-8 bytes** *(GB)*
Amazon	1	30,450	2.252	204.94	0.84
Blogs	3	18,583	0.303	85.80	0.34
Books	1	3,204	0.005	258.19	1.07
Twitter	3	19,237	2.290	62.17	0.23
Reddit	3	26,885	1.406	166.38	0.69
StackExchange	1	15,207	0.557	82.52	0.35
Total		**113,566**	**6.813**	**860.00**	**3.52**

Table 10: Subset of Large Human Language Corpus (LHLC) used as our pre-training data. The counts here are reflective original number of users and documents before going through the processing of creating multiple author-instances. Token counts are based on LLaMA 3.1 tokenizer.

A.2 Downstream Datasets Additional Details

Table 13 presents the labels proportions across downstream tasks as a result of processing downstream datasets curation similar to LHLC curation pipeline and stratifying labels distribution across train, dev, and test splits.

Dataset	Train			Dev			Test		
	Users	Docs (M)	Tokens (M)	Users	Docs (M)	Tokens (M)	Users	Docs (M)	Tokens (M)
Amazon	29,999	2.217	201.84	402	0.026	2.86	500	0.034	3.66
Blogs	20,357	0.293	83.13	780	0.013	3.91	985	0.016	4.95
Books	29,128	0.030	255.38	371	0.0004	3.01	473	0.0005	3.82
Twitter	19,266	2.194	59.72	766	0.097	2.76	976	0.123	3.52
Reddit	29,592	1.338	158.67	1,193	0.064	8.25	1,486	0.080	10.26
StackExchange	14,977	0.549	81.30	204	0.006	1.04	256	0.008	1.35
Total	**143,319**	**6.622**	**840.02**	**3,716**	**0.206**	**21.84**	**4,676**	**0.260**	**27.55**

Table 11: Train/Dev/Test split of the Large Human Language Corpus (LHLC) using recreated user statistics. Details in this table are reflective of the numbers resulting after processing the data for multiple author-instances. Token counts are based on the LLaMA 3.1 tokenizer. Number of documents and tokens are reported in millions.

Dataset	Train			Dev			Test		
	Users	Docs (M)	Tokens (M)	Users	Docs (M)	Tokens (M)	Users	Docs (M)	Tokens (M)
Amazon	29,999	2.217	201.84	402	0.026	2.86	500	0.034	3.66
Blogs	17,760	0.293	83.13	780	0.013	3.91	985	0.016	4.95
Books	2,984	0.005	255.38	194	0.0002	3.01	247	0.0003	3.82
Twitter	18,393	2.194	59.72	766	0.097	2.76	976	0.123	3.52
Reddit	25,556	1.338	158.67	1,193	0.064	8.25	1,486	0.080	10.26
StackExchange	14,977	0.549	81.30	204	0.006	1.04	256	0.008	1.35
Total	**109,669**	**6.622**	**840.02**	**3,539**	**0.206**	**21.84**	**4,450**	**0.260**	**27.55**

Table 12: Train/Dev/Test split of the subset of Large Human Language Corpus (LHLC) used for pre-training. Details in this table are based on the original number of authors and documents before processing for multiple author-instances. Token counts are based on the LLaMA 3.1 tokenizer. Number of documents and tokens are reported in millions.

Task	0	1	2	3	4
Occupation	17.2%	12.2%	21.9%	22.9%	25.9%
Movie Rev.	16.6%	15.7%	21.2%	21.82%	24.7%
Business Rev.	19.7%	23.6%	26.2%	15.0%	15.5%
Electronics Rev.	20.0%	18.5%	18.5%	20.7%	22.3%
Books Rev.	16.0%	22.0%	19.4%	21.3%	21.4%

Table 13: Approximate labels distribution percentages across downstream tasks post stratification, resulting in a relatively balanced label distributions.

A.3 Additional Analysis

Effect of author's historical context on downstream tasks. Here, we perform ablations by running each model—Llama, Llama$_{\text{LHLC}}$, and HU-Llama—in both settings, HuFT (i.e., with author's historical context) and TFT (i.e., with no human context), under the QLoRA FT setup. We find that adding the human context shows consistent benefits in the Table 16 across all downstream tasks for all models except in two specific instances alone, involving stance and sentiment detection tasks.

We report results for the classifier training setting as well in Appendix Table 15. We find trends consistent with our main results (in Table 4), where we see a general trend of performance gains with human context over no author's context for HU-Llama, and only person-level tasks benefitting in case of non-HuLM models (i.e., Llama and Llama$_{\text{LHLC}}$), with a common exception for stance and sentiment tasks across the table with no statistical significance observed.

A.4 Prompting

We use llama3.1-8b-Instruct-hf model and the vLLM framework (Kwon et al., 2023) for our prompting experiments. The experiments were run on an RTX A6000 48GB GPU. For prompting, we use 2 different methods (with and without author context). For Business and Electronics, we did additional experiments (limited author context, labeled author context). In limited author context, we limit the author context and take first [250, 500] words. Prompt details are given in Table 18.

A.5 Qualitative Analysis

We qualitatively evaluate the effect of including human context beyond the experimental analysis discussed by manually looking at downstream task predictions using a structured qualitative analysis process. We compare the predictions from the best model in the HuFT setting versus the best model in the TFT setting in two scenarios: a) HuFT predicts

		Document-Level Labels						Person-Level Labels	
		Cls ($F1$)						Cls ($F1$)	Reg (r)
Model	**FT style**	**Movie**	**Business**	**Book**	**Elec**	**Stn**	**Sent**	**Occ**	**Age**
Llama	TFT	66.40	70.93	65.32	65.26	71.37	76.44	52.45	0.882
Llama_{LHLC}	TFT	66.10	70.93	67.41	66.44^o	**73.65**	**76.60**	51.69	0.881^o
Llama	**HuFT**	67.52	**74.36**	69.22	**68.84**	69.09	76.07	52.28	**0.916**
Llama_{LHLC}	**HuFT**	**67.70**	73.56$^\wedge$	69.12	68.16$^\wedge$	72.41$^\wedge$	75.77	54.17	0.915
HU-Llama	**HuFT**	67.05*†	73.36*	**69.43***	68.22*	70.77*	75.25	**57.50***†	**0.916***

Table 14: Evaluating the effect of our LHLC data corpus on QLoRA based HuFT. Results are reported in weighted F1 for classification (cls) tasks and in pearson r for regression (reg) task. Bold indicates best in column. We indicate statistical significant ($p < .05$) difference in results between: Llama_{LHLC} HuFT and HU-Llama HuFT with †, Llama_{LHLC} TFT and HU-Llama HuFT using *, Llama_{LHLC} TFT and Llama TFT with ^{0}, and Llama_{LHLC} HuFT and Llama HUFT using$^\wedge$.

		Document-Level Labels						Person-Level Labels	
		Cls ($F1$)						Cls ($F1$)	Reg (r)
Model	**FT style**	**Movie**	**Business**	**Book**	**Elec**	**Stn**	**Sent**	**Occ**	**Age**
Llama	TFT	57.99*	65.99*	59.90*	59.06*	64.29	62.01	47.48	0.826
Llama	HuFT	55.06	63.37	57.80	57.61	62.96	61.40	53.59*	0.853*
Llama_{LHLC}	TFT	58.31*	65.64*	61.18*	60.22*	**69.25**	**63.17**	50.61	0.835
Llama_{LHLC}	HuFT	55.43	63.57	59.36	57.54	66.37	62.38	54.01*	0.855*
HU-Llama	TFT	58.99	66.15	61.15	60.15	68.21	62.72	47.38	0.839
HU-Llama	**HuFT**	**58.48**	**66.80***	**61.88***	**60.93***	67.08	62.93	**54.73***	**0.858***

Table 15: Evaluating the effect of adding author's historical language in the setting for classifier training with human context by comparing with TFT setting (i.e., no human context). Results are reported in weighted F1 for classification (cls) tasks and in pearson r for regression (reg) task. Bold indicates best in column. * indicates statistical significance $p < .05$ for each model separately between their respective TFT and HuFT results, with * marked on the better result between the two.

		Document-Level Labels						Person-Level Labels	
		Cls ($F1$)						Cls ($F1$)	Reg (r)
Model	**FT style**	**Movie**	**Business**	**Book**	**Elec**	**Stn**	**Sent**	**Occ**	**Age**
Llama	TFT	66.40	70.93	65.32	65.26	71.37	76.44	52.45	0.882
Llama	**HuFT**	67.52*	**74.36***	69.22*	**68.84***	69.09	76.07	52.28	**0.916***
Llama_{LHLC}	TFT	66.10	70.93	67.41	66.44	**73.65***	76.60	51.69	0.881
Llama_{LHLC}	**HuFT**	**67.70***	73.56*	69.12*	68.16*	72.41	75.77	54.17	0.915*
HU-Llama	TFT	66.19	70.82	67.24	65.85	72.32	**78.17***	52.02	0.884
HU-Llama	**HuFT**	67.05*	73.36*	**69.43***	68.22*	70.77	75.25	**57.50***	**0.916***

Table 16: Evaluating the effect of adding author's historical language in the QLoRA setting with human context (HuFT) by comparing with QLoRA TFT setting (i.e., no human context). Results are reported in weighted F1 for classification (cls) tasks and in pearson r for regression (reg) task. Bold indicates best in column. * indicates statistical significance $p < .05$ for each model separately between their respective TFT and HuFT results, with * marked on the better result between the two.

correctly and TFT predicts incorrectly, and b) vice-versa. Next, we sort the instances by the amount of author's historical context used in these cases and manually review to find where this human context proved to be helpful (i.e., first scenario), and where the human context may turn out to be ambiguous and thus hurting the model's judgement. Additional qualitative examples are presented in Appendix Tables 19, 20, 21, 22, 23, 24, 25, and 26.

Model	Human Context (HC)	Document-Level Labels Cls ($F1$) Business	Elec
Llama-Instruct	No HC	59.17	59.72
Llama-Instruct	Full HC	60.34	**61.16**
Llama-Instruct	Limited HC (500 words)	60.29	**61.17**
Llama-Instruct	Limited HC (250 words)	**60.95***	61.10
Llama-Instruct	Labeled HC	60.63^{o}	56.44^{o†}

Table 17: Results with ablations for directly including human context (HC) in prompting using Llama-Instruct 3.1 8B model over selected downstream tasks. Results are reported in weighted F1 for classification (cls) tasks. Bold indicates best in column. We indicates statistical significance ($p < .05$) difference in results between: best in column and Full HC using *, Labeled HC and No HC with ^{o}, and Labeled HC and best in column using †.

Task	Prompt Template
Stance Topic	Identify the stance of the given target text towards {topic}. Select one of the three: In Favor, or Against, or Neutral. Here is the target text: {text} Do not include any extra information.
Stance Topic with Author Context	Here is a list of the previous messages written by the person in chronological order to learn more about the person: {messages} Identify the stance of the given target text towards {topic}. Select one of the three: In Favor, or Against, or Neutral. Here is the target text: {text} Do not include any extra information.
Sentiment	Identify the sentiment of the given target text. Select one of the three: Positive, or Negative, or Neutral. Here is the target text: {text} Do not include any extra information.
Sentiment with Author Context	Here is a list of the previous messages written by the person in chronological order to learn more about the person: {messages} Identify the sentiment of the given target text. Select one of the three: Positive, or Negative, or Neutral. Here is the target text: {text} Do not include any extra information.
Reviews Classification	Determine the star rating (from 1 to 5) that best reflects the following {review} and answer strictly as only one from [1,2,3,4,5]. {review}: {text} Do not include any extra information.
Reviews Classification with Author Context	Here is a list of the previous {category} written by the person in chronological order to learn more about the person: {messages} Determine the star rating (from 1 to 5) that best reflects the following {review} and answer strictly as only one from [1,2,3,4,5]. {review}: {text} Do not include any extra information.
Reviews Classification with labeled Author Context	Here is a list of the previous {category} written by the person in chronological order and the corresponding rating that they gave: {review rating pairs} Determine the star rating (from 1 to 5) that best reflects the following {review} and answer strictly as only one from [1,2,3,4,5]. {review}: {text} Do not include any extra information.
Job Classification	Given a list of messages written by a person, predict their most relevant job category as only one of the following: Education, Student, Technology, Arts, Communications-Media. Here is a list of the person's written messages in chronological order: {messages} Now predict the person's job category. Do not include any extra information.
Age Estimation	Given a list of messages written by a person, estimate the person's age. Here is a list of the person's written messages in chronological order: {messages} Now just give the person's age as a real valued number without any explanation. Give only the age value between 0 to 100 and no other text.

Table 18: Prompting Templates, where topic = [Hillary Clinton, atheism, feminism, legalization of abortion, climate change as a real concern], review = [Movie review for Movies, Business review for Business, review for books and electronics], category = [movie reviews, reviews for different businesses, reviews for books available on amazon, reviews for electronic products available on amazon]. {review rating pairs}, {messages} are separated by a line. For "limited author context" we limit the {messages} to [250, 500] words.

Task	Target Text	Gold	With HC	No HC	Selected Historical Language	Why History Helps/Hurts
Stance Pred. (Abortion)	@_msthirdward @BuhayIpaglaban @blueskies366 It's never been about "life" has it? "#Prolife" are OK with women dying.	AGAINST	AGAINST	FOR	[...] It's never been about "life" has it? "#Prolife" are OK with women dying. [...] The culture of death is #prolife who deny pregnant people healthcare. [...] thank you again to show that you are AGAINST CONTRACEPTION; "#Prolife" = a code word for misogyny. [...] Tell me more about how #antichoice laws kill 47,000 women A YEAR! [...] Remember Savita, killed by Irish "#prolife" laws? Repeal the 8th. [...]	Without historical context, the model interprets the tweet as ambiguous or rhetorical. User history contains strong, repeated anti-prolife and pro-choice language, which disambiguates stance.
Stance Pred. (Atheism)	I #DenounceHarper for refusing to include family planning in foreign aid even though spending $1 could save $6 #wherestheFP	FOR	FOR	AGAINST	Enough sneaking around, bringing in religious groups into public schools to access our kids in secrecy. [...] I wonder if anti-choice advocates denying accurate SexED ever consider the harm they cause. [...] How many can you fit in your womb — mocking reproductive control arguments. [...] Anti-choice actors never try to police men over sex, but punishing women puts them into a frenzy.	The target criticizes a political decision but does not explicitly mention religion. History shows persistent opposition to religious influence in policy/education, enabling the correct inference.
Stance Pred. (Abortion)	I #DenounceHarper for refusing to include family planning in foreign aid even though spending $1 could save $6 #wherestheFP	FOR	FOR	AGAINST	family planning as part of foreign aid. [...] Sneaking religious groups into public schools to access kids in secrecy. [...] Pregnant teens denied accurate SexED suffer long-term consequences. [...] Talking about fertility clinics and rescue rhetoric highlights anti-choice framing around women's bodies and wombs. [...]	Without context, the model associates foreign-aid criticism with conservative framing. History consistently links family planning/contraception with women's health and anti-restriction advocacy.

Table 19: Stance prediction qualitative examples where incorporating human context (HC) changes the model's interpretation and improves stance inference.

Task	Target Text	Gold	With HC	No HC	Selected Historical Language	Why History Helps/Hurts
Stance Pred. (Feminism)	So if a women earns money then it's HERS to KEEP?? #feminismiscruelty	AGAINST	FOR	AGAINST	[...] So if a woman earns money then it's HERS to KEEP?? [...] UGH I am SICK to DEATH of this whole women wanting "equality" thing. [...] Drawing attention to women's health issues is framed as evidence that feminism is harmful or cruel. [...] A woman wanting to be equal to a man is portrayed as something outrageous and unnatural. [...] Feminism is characterized as a hate group rather than an equality movement. [...]	History is repetitive and emotionally charged. The model over-indexes on sarcasm/hostile tone, which can flip polarity despite a consistent anti-feminist intent.
Stance Pred. (Abortion)	A prochoice advocate but circumcise ur baby? Fucking hypocrite! #circumcision #humanrights	AGAINST	FOR	AGAINST	[...] A pro-choice advocate framed as morally hypocritical for supporting circumcision. [...] If a man wants abortion but the woman wants to keep it, he should not be responsible for child support. [...] Modern feminism is described as irrational and self-serving. [...] Violence against women is contrasted with claims that feminism focuses on trivial cultural issues. [...] Broader grievances about gender roles dominate the historical context. [...]	History spans multiple issues (abortion, feminism, gender roles), causing topic drift. The model generalizes from broader ideology instead of anchoring to the abortion-specific claim.

Table 20: Stance prediction qualitative examples where incorporating human context (HC) does not help: history can amplify tone or introduce topic drift, leading to incorrect predictions.

Task	Target Text	Gold	With HC	No HC	Selected Historical Language	Why History Helps/Hurts
Review Pred. (Amazon)	If you have a hard time falling asleep, buy this DVD — you won't regret it. You'll be asleep before it ends. The only way to see how it ends is to fast forward to the last 10 minutes and you will still not know what the plot is about. Even if you manage to keep your eyes open throughout the entire DVD.	0.0	0.0	4.0	[. . .] I did not care for this movie . . . reading cost nothing but your time and eyesight . . . the DVD does not disclose anything more than you can find out for yourself! (0.0) [. . .] the movie is extremely boring, every scene is predictable . . . Don't waste your time like I did. (0.0) [. . .] this is a very poorly written movie . . . the preview makes you believe this might be funny — it is far from being funny. (0.0) [. . .] The most entertaining movie I have seen in a very long time . . . It is a must see movie. (4.0) [. . .] The script was well written, the acting superb . . . one of the funniest movies I've seen in a long time. (4.0) [. . .]	Without historical context, the sarcastic "you won't regret it" reads as genuine praise. The user's history shows a consistent pattern where sleep/boredom metaphors co-occur with "extremely boring" and "Don't waste your time", clarifying this as a negative review.
Review Pred. (Amazon)	What a horrifying movie! I wish Tim would have gotten an Oscar for this movie.	4.0	4.0	0.0	[. . .] What an all-star cast that all performed excellent! The ending was great as well. Sean Penn did an incredible job on this film. (3.0) [. . .] The script was well written, the acting superb; this movie is one of the funniest I've seen in a long time. (4.0) [. . .] This and Revenge have been Kevin Costner's best performances. Amazing script, directing, and acting. (4.0) [. . .] By contrast, clearly negative reviews contain unambiguous dismissal such as "extremely boring", "don't waste your time", and "very poorly written", all paired with low ratings (0–1). [. . .]	Without historical context, "horrifying" is treated as a negative cue. The user's history makes clear that strong sentiment words are not decisive, while consistent emphasis on performance/acting praise aligns with high ratings.

Table 21: Amazon reviews qualitative examples (with vs. without human context). Positive examples where human context helps.

Task	Target Text	Gold	With HC	No HC	Selected Historical Language	Why History Helps/Hurts
Review Pred. (Amazon)	I have DVD's from other region zones and PAL DVD's and they all play on my Blu Ray... why wouldn't this one? Someone who has gotten this and is playing it on a US Blu Ray player, please respond!	4.0	0.0	4.0	[...] I had the soundtrack on Colgems records and loved the music and dialog; when I saw the movie was on DVD, I grabbed it! Great nostalgia. (3.0) [...] I read people complain that this is long and boring, but this documentary is like no other in rock history. Morrison was a genius, complicated and tortured, and this film opened my eyes. (4.0) [...]	The review is a neutral, practical question about disc compatibility, not a complaint. Without history, the model treats it as informational and predicts a positive rating. With history, the model expects explicit enthusiasm and overcorrects, interpreting the restrained tone as dissatisfaction.
Review Pred. (Amazon)	I purchased this new from media_distributors but I made the mistake of not looking at it right away. It looked like it was new and the video was labeled Blonde Crazy... When I played it, it turned out to be Rumpole. I wrote an email to Media Distributors after posting this, and they replaced my video. Outstanding.	0.0	4.0	0.0	[...] A collection of way overpriced videos that you can get in bargain bins for $5 each. (1.0) [...] The Alpha version is of very poor video quality, though the underlying film may have been interesting. (1.0) [...] This is not a great film, but it is a fun film and interesting to see Robert Mitchum in his first feature. (2.0) [...] The Alpha version of Millie is exceptional with outstanding print quality. (3.0) [...]	The review reports a shipping error that was resolved, but the overall experience is still framed around a mistake. Without history, the model anchors on the problem-focused content and predicts a low rating. With history, prior instances of praising releases/distributors after logistics make the model overweight a single positive cue ("Outstanding") and inflate the rating.

Table 22: Amazon reviews qualitative examples (with vs. without human context). Negative examples where human context does not help.

Task	Target Text	Gold	With HC	No HC	Selected Historical Language	Why History Helps/Hurts
Review Pred. (Yelp)	Do you have a heart condition, high cholesterol, diabetes, overweight, or anything like that? Well, here is your swan song that's worth the further damage — the Daily Dog at Lady Di's. A deep fried hot dog stuffed with gooey cheese, wrapped in crispy bacon. Eat one, you'll find nirvana. Eat two, you'll find the ER.	4.0	4.0	0.0	[...] Oh dear laaawd! That cheeseburger on a pretzel bun with onion straws — are you kidding me?!? (3.0) [...] Fast food is standard everywhere, but this Taco Bell always gets the order wrong. Mexican pizza with tomatoes and no cheese instead. (1.0) [...] Chuck E Cheese has mediocre pizza, dirty salad bars, and chaos everywhere. This place has it all — and not in a good way. (0.0) [...]	In isolation, the exaggerated health warnings ("ER", "damage") read as criticism. The user's history shows that this kind of sarcastic, over-the-top storytelling is how they express strong approval. With context, the humor and sensory detail signal a positive experience rather than a complaint.
Review Pred. (Yelp)	Hey Megabus, guess where I am right now? Yeah, that's right — I'm on Bolt Bus writing a review because Bolt Bus actually gave me the wifi they promised me. Megabus, you are always Megalate, Megarude, and I will not be taking you anymore.	0.0	0.0	4.0	[...] Their service is terrible! I will NEVER recommend them to anyone. (0.0) [...] I'm giving this place 3 stars because of their horrible service. The bagels are great, but the service leaves a bad taste. (2.0) [...] I really wanted to enjoy the food and service, but they fell short in both. (1.0) [...] This truly is the best pizza ever! Everything tastes fresh and the staff are inviting. (4.0) [...]	Without historical context, the review reads as enthusiastic because it highlights concrete benefits (wifi) and praises a competitor, leading to a positive prediction. In the user's history, genuine positives are direct and unambiguous, while sarcastic, insult-heavy rants like this one consistently map to 0-star ratings, disambiguating the intent as negative.

Table 23: Yelp reviews qualitative examples (with vs. without human context). Positive examples where human context helps. For these cases, red highlights in the target indicate misleading cues without context, while green highlights in history indicate disambiguating signals.

Task	Target Text	Gold	With HC	No HC	Selected Historical Language	Why History Helps/Hurts
Review Pred. (Yelp)	I've never had Chinese food before, and thought to myself after the first bite, "wow this tastes like it was just cooked for me." This is the experience we had not once but every time we have ordered from Lingnan.	4.0	1.0	4.0	[. . .] The first thing that made me decide I would never eat here was the smell and the condition of their kitchen. I was gagging. (0.0) [. . .] We were appalled to find they didn't offer refried beans. We never ended up eating anything. (1.0) [. . .] This place must be stuck in the early 90's. No debit, overpriced ATM — no thanks. (1.0) [. . .] Despite the flies and pricing, the pulled pork sandwich was great, but a lot really pissed me off. (2.0) [. . .]	The review is clearly positive, with direct praise and repeated satisfaction. Without history, the model follows this surface sentiment. With history, frequent complaint-heavy reviews set a low baseline, causing the model to discount brief praise and predict too low.
Review Pred. (Yelp)	I love this place for some late night eats. I would drive past Viva Burrito and go to El Potosino because they have bigger and better burritos. Viva is dirty and the employees are lazy, but El Potosino has great California burritos with guacamole and is one of the best late night eats for the price, quantity, and quality.	3.0	1.0	3.0	[. . .] I don't understand why people still come to this restaurant — the food has gone downhill and is overpriced. (1.0) [. . .] Fast and ok, but the new items were soggy and I wouldn't recommend them. (2.0) [. . .] This place is terrible. Bad service, bad food — I wouldn't recommend it at all. (0.0) [. . .] I do love this place, but the service is terrible and there's no urgency. (0.0) [. . .]	The user criticizes one restaurant but is clearly endorsing another. Without history, the model follows that intent. With history, frequent past complaints make the model overreact to negative words and overlook the recommendation.

Table 24: Yelp reviews qualitative examples (with vs. without human context). Negative examples where human context does not help. For these cases, red highlights in history indicate misleading priors, while green highlights in the target indicate the intended sentiment.

Task	Target Text	Gold	With HC	No HC	Selected Historical Language	Why History Helps/Hurts
Occ	"People talk about all this unused potential in the world, but the greatest untapped power is the human mind…" [TFT → Arts] [...] "Language is fascinating — the brain doesn't process words the way we think it does." [TFT → Arts] [...] "If we limit what ideas people are exposed to, we limit the future itself." [TFT → Arts] [...] "... homeward bound :: need my hermit shell back ::" [TFT → Technology]	Tech	Tech	Arts ["Arts", 35, "Tech.", 24, "Student", 13, "Comm.-Media", 6, "Edu", 4]	i am mack, i will some day become great [. . .] People talk about unused CPU power [. . .] the human brain [. . .] Discussion of banning evolution from science classes [. . .] Reactions to NASA discoveries and scientific progress [. . .] homeward bound :: need my hermit shell back [. . .]	Without history, expressive and philosophical phrasing is easy to read as artistic or abstract writing. Across the user's history, repeated references to computing metaphors, science education, and technology–policy reasoning clarify a technology-oriented mindset rather than an arts identity.
Occ	"Played Doom III with Bob yesterday from 2:00 until dark. At one point I noticed strange twinkling lights coming from the bottom of his blinds. Turns out it was the sun. We had no idea what time it was." [TFT → Tech.] [...] "Now I'm off to Future Shop. The Zoom button on my camera has become stuck, and I need it repaired. Wish me luck." [TFT → Tech.] [...] "The last few days of school annoy me. Monday and Tuesday are the last official days, but attendance is extremely sporadic — people would much rather study for exams or play in the sun than sit in classes. Besides, the teachers have all but given up teaching; classes consist of movies, maybe a pretest or two." [TFT → Student] [...]	Student	Student	Tech ["Tech.", 25, "Student", 17, "Edu.", 17, "Arts", 14, "Comm.-Media", 10]	Last official day of school [. . .] had a chemistry final [. . .] Success in my math final and biology provincial [. . .] Currently very pleased with Future Shop [. . .] Classes are completely done [. . .] Exams start tomorrow [. . .] play some Super Nintendo [. . .]	Without broader history, isolated mentions of tools or hobbies push the prediction toward Technology. The full history shows a consistent focus on classes, exams, and academic routines, establishing a clear student-centered identity.

Table 25: Person-level job prediction examples (with vs. without human context). Target text is omitted for user-level classification; history provides the primary signal.

Task	Target Text	Gold	With HC	No HC	Selected Historical Language	Why History Helps/Hurts
Occ.		Tech.	Student	Tech.	study F.6 … subjects choices are: Physics, Biology / Pure Mathematics, Computer Studies … A-level syllabus is very demanding [...] release of HKCEE result … procedure for applying F.6 [...] had a chemistry final … success in my math final [...] decided to write a polling web page using ASP.NET … learned stored procedures for SQL Server [...] downloaded Visual C# Express … mixing .NET and native code [...]	The history mixes strong school-centered signals (exams, subject selection, results) with clear hands-on technical work. When history is used, the model overweights repeated school references and collapses the user into a Student identity. Without history, it relies more on concrete development activity (ASP.NET, C#, SQL), aligning better with Tech..
Occ.		Student	Tech.	Student	3rd party web browser for any modded Xbox [...] HTML 4.0 support, HTTP 1.1, Javascript support [...] prefer Mozilla/Firefox within Xbox Linux [...] stand-alone app launchable from any dashboard [...] Xbox modding tutorials [...] release date of Microsoft Windows XP Service Pack 2 [...] played around with technical betas [...] security dashboard, spyware, pop-ups [...] Ok so you wana play HALO xbox online [...] step-by-step setup instructions [...] simple and easy to use, free app [...] learning-oriented walkthrough [...] Really cool forum of a friend of mine [...] perfect if you're into halo [...] sharing screenshots and experiments [...] curiosity-driven tinkering rather than deployment [...]	The history contains many advanced technical terms and tools, which makes the user appear professionally technology-oriented. However, these posts are mostly framed around learning, gaming, and experimentation, with tutorial-style explanations and hobbyist motivation. Using history amplifies the technical surface cues and misleads the prediction away from the user's true role as a Student.

Table 26: Negative examples where adding human context (HC) can mislead predictions for person-level job predictions. Target text is intentionally left blank.