Addressing the Ecological Fallacy in Larger LLMs with the Author's Context

Anonymous ACL submission

Abstract

Language model training and inference ignores a fundamental fact about language– the dependence between sequences of text that come from the same person. Prior work had shown that addressing this form of *ecological fallacy* can greatly improve performance of a smaller language model, a 110 M parameter GPT-2 model. 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 which 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. Empirical comparisons show that addressing the ecological fallacy during fine-tuning alone improves the performance of the larger 8B model over standard fine-tuning, as well as prompting with an instruction-tuned variant. 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, a so-called *ecological fallacy* of assuming sequences from the same person are independent (or treated the same as those from different people). This leaves models with imperfect representations of the world and little means to directly address biases (Soni et al., 2024) as they consistently lack variance in their expressed psychological traits (Varadarajan et al., 2025).

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 author's context as modeled by their previous texts. Recent work has suggested that addressing the ecological fallacy during continued pretraining by turning the LM task into *human language modeling* (HuLM) can improve performance both in terms of LM perplexity and downstream applicaitons for a small scale model (a variant of the 124M parameter GPT-2 model) (Soni et al., 2022). Language was modeled during pretraining in the context of the temporally ordered texts from the same user (we refer to this as the HuLM context).

However, it is not clear *a priori* that the more powerful larger models need this additional author HuLM context. One may posit that LLMs with billions of parameters trained over trillions of tokens, capture language from a large population of humans and may overcome any representational or distributional shortcomings that arise from lack of processing in the author's context.

To address this, we explore different ways of incorporating author contexts into 8B sized models derived from the open source Llama weights. First, we consider continued pre-training of the Llama weights the HuLM objective from (Soni et al., 2022). This adds the author HuLM context to pretraining. Second, we also setup tasks where we inject the author HuLM context in fine-tuning. For each task instance, we also have the author's history which we process into the HuLM context of temporally ordered message sequence. We compare the performance of these models against other standard ways of using Llama 8B models (both in fine-tuning and prompting).

Our empirical results highlight several important findings: (i) Modeling language in the author's context provides substantial improvements on multiple tasks. (ii) HuLM style pretraining doesn't

add gains when compared to directly fine-tuning Llama-8B with HuLM author contexts. (iii) For most tasks fine-tuning with HuLM author contexts outperforms simply prompting (instruction-tuned) Llama with history. These results clearly show that addressing the ecological fallacy clearly benefits even larger 8B scale 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. 2) developed a bigger HuLM model (in the range of 8B parameters) and 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 stack exchange (Lambert et al., 2023). 3) expanded tasks and dataset with author context via author's historical texts.

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 or their demographics in multiple applications, such as sentiment analysis (Mireshghallah et al., 2021) or reducing social biases (Garimella et al., 2022). In the realm of large LLMs, prior work has shown benefits in considering a person's dynamic emotional states 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).

Recent works (Soni et al., 2022) suggest including the author's context within the pre-training task of next word prediction. 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. They build a Human-aware Recurrent Transformer (HaRT) pre-trained for the HuLM task. We briefly describe HaRT's architecture here as out work builds on the concept of HuLM and assessing the impact of processing language within the author's context in larger LLMs.

HaRT is an autoregressive model consisting of 12-layers initialized with GPT-2 small weights. It modifies GPT-2's architecture to use a *user state* U at an initial layer (layer 2) in the self-attention computation. This U is concatenated to the hidden states from the first layer and transformed to create the query vector in layer 2. Additionally, the user state is recurrently updated by adding the previous user state to the hidden states from a later layer (layer 11) using transformations and *tanh* activation. Essentially, the model latently learns the user state by recurrently processing long contexts of temporally-ordered texts written by the same author.

At the same time, larger LLMs have demonstrated remarkable performances in many tasks (OpenAI, 2023; Hendrycks et al., 2021; Jimenez et al., 2023; Singhal et al., 2022). However, larger LLMs are not yet evaluated for the effects of processing language within the author's context when continued to pre-train or when fine-tuned for downstream tasks. In this study, we extend the concept of HuLM into larger LLMs in terms of continued QLoRA pre-training and fine-tuning for downstream tasks.

3 Models and Methods

We seek to evaluate the effectiveness of processing language within the author's context, i.e., mitigating the *ecological fallacy*, in larger LLMs. This can be assessed at different levels: pre-training, fine-tuning, and prompting. Prior works have built human language models at smaller scales (Soni et al., 2022) and fine-tuned them for downstream tasks by collectively processing language written by the same author (i.e., within the author's context), or experimented with user-centric prompting (Salemi et al., 2023). However, the need to evaluate larger LLMs for fine-tuning within the author's context or pre-training for the HuLM task remains. Here, we consider continued pre-training for the HuLM task with QLoRA (Dettmers et al., 2023) for larger LLMs, fine-tuning for downstream tasks with and without the author's context, and briefly comparing prompting for downstream tasks within the author's context for completeness.

To this end, we select Llama 3.1 8B (Grattafiori et al., 2024) as our base model. It adopts a decoder-only transformer architecture, enabling us to adapt it to the autoregressive HuLM task easily. Here, we build 2 variants of bigger HuLM models: HU-Llama and HaRT$_{\text{Llama}}$, following the data processing and HaRT architecture from Soni et al. (2022) re-

Dataset	Epochs	Users	Docs *(millions)*	Tokens *(millions)*	UTF-8 bytes *(GB)*
Amazon	1	30,902	2.276	208.37	0.86
Blogs	3	19,525	0.322	91.99	0.36
Books	1	3,425	0.005	262.21	1.09
Twitter	3	20,135	2.414	66.00	0.24
Reddit	3	28,229	1.482	177.15	0.73
StackExchange	1	15,440	0.564	83.69	0.35
Total		**117,656**	**7.063**	**889.41**	**3.64**

Table 1: Subset of Large Human Language Corpus (LHLM) used as our pre-training data. Token counts are based on LLaMA 3.1 tokenizer.

spectively. Further, we assess the impact of human-level fine-tuning (HuFT), i.e., fine-tuning for downstream tasks by processing text documents within the author's context, over traditional fine-tuning, i.e., fine-tuning for downstream tasks by *independently* processing text documents written by the same author.

We describe the models and methods we use below. More details on training and fine-tuning processes can be found in section 5.

HU-Llama. We continue to pre-train Llama for the next word prediction task using the LHLC (see Section 4.1), however, we do so over temporally-ordered documents written by a particular author, concatenated using a special *insep* token, instead of randomly sampling documents and processing them independently. This results in each instance representing an author, inducing an explicit author's context by introducing dependence on language written by the same author.

HaRT$_{\text{Llama}}$. Although Llama efficiently handles long contexts (i.e., 8192 tokens), we seek to scale HaRT's recurrence architecture (Soni et al., 2022) (see section 2 for a brief overview) into larger LLMs to compare with smaller-scale recurrent HuLM models. Thus, we implement a similar recurrent architecture HaRT$_{\text{Llama}}$ using user states at an initial layer (layer 2) and updating the user states using hidden states from a later layer (layer 31).

Llama, Llama$_{\text{LHLC}}$, HaRT$_{\text{GPT2-twt}}$. We compare the bigger HuLM models with their counterpart non-HuLM models: Llama and Llama$_{\text{LHLC}}$. We adapt Llama to our pre-training data corpus (LHLC) and call it Llama$_{\text{LHLC}}$. Additionally, we fine-tune the publicly available HaRT model (we call it HaRT$_{\text{GPT2-twt}}$) for our downstream tasks as well as report published numbers from the paper (we call it HaRT$_{\text{GPT2-twt-fb}}$) (Soni et al., 2022).

HuFT: Human-level Fine-tuning. We compare the performance of fine-tuning HU-Llama and Llama$_{\text{LHLC}}$ models for downstream tasks in two ways: with no author's context (traditional FT), and with author's context (HuFT). More details on experimenting with traditional fine-tuning and HuFT can be found in sections 5.3 and 5.2.

Prompting within Author's Context. We compare the results on downstream tasks against prompting the Llama 3.1 8B instruction-tuned model in two ways: input one document at a time, and inducing the author's context by providing a list of temporally-ordered documents written by a particular author (details in Appendix A).

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 150K authors, and a data report containing the details of our dataset construction and design principles [REDACTED]. Briefly, LHLC creation steps include: 1) removing missing data, 2) data deduplication, 3) english filtering, 4) text formatting (encoding, URLs, etc.), 5) toxicity filtering, 6) anonymization. Here, we use a subset of this data summarized in Table 1.

Task	Users	Docs	Labels	med #wpd	max #wpd	med #dpa	max #dpa	med #wpa	max #wpa
Age$_{\text{blogs}}$	15,942	210,253	15,942	100	350	9	353	1189	4999
Age$_{\text{wassa}}$	729	3,264	729	72	166	5	10	350	1256
Occupation	3,539	47,500	3,539	95	350	8	337	1166	4999

Table 2: We curate three person-level task datasets using documents from the blogs and WASSA essays corpora. We apply an inclusion criteria resulting in the above statistics. Here, wpd = words per document, dpa = documents per author, wpa = words per author. Additionally, we have a minimum of 10 words per document in blogs and 40 in WASSA essays, and a minimum of 250 words per author in blogs and 50 words per author in WASSA essays.

Task	Train			Dev			Test		
	Users	Docs	Labels	Users	Docs	Labels	Users	Docs	Labels
Age$_{\text{blogs}}$	10,354	135,594	10,354	2,402	32,773	2,402	3,186	41,886	3,186
Age$_{\text{wassa}}$	434	1,961	434	75	346	75	220	957	220
Sentiment	6,246	28,808	6,461	1,000	4,548	1,030	2,859	13,924	2,994
Stance	1,361	11,318	1,658	332	1,996	418	768	4,097	945
Occupation	2,135	28,199	2,135	532	7,770	532	872	11,531	872

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.

4.2 Downstream Datasets and Tasks

We evaluate the effectiveness of human language modeling and human-level fine-tuning on five downstream tasks. These tasks can be categorized into two types: **document-level** and **person-level**. In the first type, given a target text sequence (document) written by a person, the model is required to predict the label (e.g., stance of a person on a topic like *atheism*). In the second type, given multiple text sequences (documents) written by a person, the model is required to predict/estimate the label (e.g., the occupation or age of the person).

Document-Level. We evaluate the HuLM and HuFT models on the 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.

Person-Level. Similar to LHLC, we curate three person-level downstream task datasets from existing sources—blogs (Schler et al., 2006) and WASSA essays (Barriere et al., 2022, 2023; Giorgi et al., 2024c)—that we will release publicly. We evaluate HuFT on the tasks of classifying a person's **occupation** from the blogs written by them, and estimating a person's **age** from the blogs written by them (**Age**$_{\text{blogs}}$) or from the essays written by them (**Age**$_{\text{wassa}}$). We limit the occupation classification data to consist of the top 5 occupations from blogs corpus (student, technology, arts, communication and media, and education). We clean, process, and apply inclusion criteria to the datasets and create train, validation, and test splits. Details on stats and splits can be found in Tables 2 and 3.

5 Training and Experiments

This section describes the training framework for building bigger HuLM models, and the fine-tuning and HuFT processes on downstream tasks. We build on code from the Hugging Face (HF) library (Wolf et al., 2020).

5.1 HuLM Pre-training

Here, each input instance represents an author's language, where text documents written by the same author are concatenated using the special *insep* token as the separator in a temporal order. HU-Llama caps the context length at 8192 tokens for each author-level instance. For consistency, HaRT$_{\text{Llama}}$ caps the maximum blocks to 8 with a block size of 1024, resulting in 8192 tokens per author-level input instance. Similarly, we adapt Llama$_{\text{LHLC}}$ by continuing pre-training at the document-level using ditto tokens, i.e., each document is processed independently.

We train these models using low-rank adapters and 4-bit quantized weights (QLoRA) (Dettmers et al., 2023) in a distributed environment using Ac-

Model	Document-Level		Person-Level		
	Stance ($F1$)	Sentiment ($F1$)	Occupation ($F1$)	Age_blogs (r)	Age_wassa (r)
Llama	56.52	**78.19**	54.28	0.887	0.543
Llama$_{LHLC}$	57.83	77.87	53.74	0.887	0.547
HuFT-Llama$_{LHLC}$	**66.55***	77.27	**57.07****	**0.919**	**0.617***

Table 4: Role of HuFT, Human-level Fine-Tuning, for downstream tasks: We find substantial gains by processing language in the author's context as modeled by their previously written texts. Results are reported in weighted F1 for Stance, Sentiment, and Occupation classification, and pearson r for estimating age. Bold indicates best in column and * indicates statistical significance with $p < .0001$ and ** with $p < 0.05$. We use a permutation test for classification tasks and paired t-test for regression tasks for HuFT-Llama$_{LHLC}$ against Llama$_{LHLC}$.

Model	Document-Level		Person-Level		
	Stance ($F1$)	Sentiment ($F1$)	Occupation ($F1$)	Age_blogs (r)	Age_wassa (r)
HaRT$_{GPT2\text{-}twt}$	70.48*	74.31	42.80	0.710	0.245
HaRT$_{GPT2\text{-}twt\text{-}fb}$ (Soni et al., 2022)	**71.1**	**78.25**	-	-	-
HaRT$_{Llama}$	63.10	49.49	49.23	0.908	0.327
HU-Llama	67.39	76.98	**57.69**	0.916	0.592
HuFT-Llama$_{LHLC}$	66.55	77.27	57.07	**0.919***	**0.617***

Table 5: Role of scale of LLMs in HuLM models: We find simply HuFT-Llama$_{LHLC}$ to be similar to HU-Llama fine-tuned for the downstream tasks. Additionally, the recurrent HuLM architecture (HaRT) does not appear to scale for larger LLMs. However, this may be attributed to various factors such as training only a few model parameters due to compute limitations. Interestingly, smaller-scale HuLM models demonstrate better performance on document-level tasks as compared to larger-scale HuLM models. Bold indicates best in column and * indicates statistical significance with $p < .0001$.

celerate (Gugger et al., 2022), accommodating for compute availability. Additionally, we use mixed-precision training, performing operations in half-precision format to speed up computation. We use the PEFT library (Han et al., 2024) integrated in HF to train weights associated with Q, K, V, O, and additionally for HaRT$_{Llama}$, weights associated with modified Q in layer 2 and the recurrent user states module weights associated with U (user states) and hidden states H from layer 31. At the 8B scale, this setup enables us to continue HuLM pre-training with a batch size of 3 per GPU on our hardware and 8192 tokens per author-instance. Similarly, we use a batch size of 123 with each instance (representing each document) limited to 200 tokens for Llama$_{LHLC}$ continued pre-training. We run initial experiments with smaller data samples using learning rates 1e-6, 3e-4, and 5e-5, and resort to 5e-5 when training with full data. We train on full data incrementally, and the details on the number of epochs each data source was trained can be found in Table 1.

5.2 HuFT: Human-level Fine-tuning for Downstream Tasks

Here, similar to pre-training input instances, we concatenate temporally-ordered documents written by the same author, separated by the special *insep* token. For document-level tasks, the models process all the tokens in the concatenated sequence and use the target document's last token's hidden states (from the last layer) to predict the author's stance or sentiment. For person-level tasks, the models process the author's language similarly and use the averaged token embeddings across all tokens from an author to predict/estimate the person's occupation or age. This way of HuFT allows mitigating the ecological fallacy in larger LLMs at the fine-tuning level.

While HuLM models are designed to adopt HuFT, standard LLMs are not known to use this approach, usually due to the limitation of context lengths. However, larger LLMs that can process larger contexts have not been evaluated for HuFT. So in addition to the HuLM models (HU-Llama, HaRT$_{Llama}$, HaRT$_{GPT2\text{-}twt}$), we also evaluate Llama$_{LHLC}$ using HuFT for downstream tasks, leveraging its capacity to process long contexts.

Model	Document-Level		Person-Level		
	Stance ($F1$)	Sentiment ($F1$)	Occupation ($F1$)	Age_blogs (r)	Age_wassa (r)
Llama	56.52	**78.19**	54.28	0.887	0.543
Llama$_{LHLC}$	57.83	77.87	53.74	0.887	0.547
HU-Llama $_{4096\ No\text{-}HuFT}$	57.56	**78.13**	54.02	0.887	0.550
HU-Llama $_{4096}$	**67.39***	76.98	**57.69****	**0.916**	**0.592***
HU-Llama $_{8192}$	-	-	57.26	**0.919***	-

Table 6: Role of the amount of author's context on HuLM models downstream task performances: We find these results to show that using the author's context helps across the board. Additionally, showing plateauing performance when increasing the amount of author's context further. Bold indicates best in column and * indicates statistical significance with $p < .0001$ and ** with $p < 0.05$.

Model	Document-Level		Person-Level		
	Stance ($F1$)	Sentiment ($F1$)	Occupation ($F1$)	Age_blogs (r)	Age_wassa (r)
Llama prompt	68.44	65.75	-	-	-
Llama prompt within Author's context	**69.92***	64.91	37.09	0.145	0.246
Llama	56.52	**78.19**	54.28	0.887	0.543
Llama$_{LHLC}$	57.83	77.87	53.74	0.887	0.547
HuFT-Llama$_{LHLC}$	66.55	77.27	**57.07****	**0.919**	**0.617***

Table 7: We find prompting to be insufficient to understand language in the author's context for person-level tasks. At the same time, we see gains in stace detection. Bold indicates best in column and * indicates statistical significance with $p < .0001$ and ** with $p < 0.05$.

Topic	Tweet	Actual Stance	Stance with HuFT	Stance without HuFT	Selected Author's Context	Word Cloud from Author's Context
Clinton	The federal government did not create the states, the states created the federal government. Ronald Reagan #Federalism	NEGATIVE	NEGATIVE	NEUTRAL	• It's sad that states can force you to join an organization. What happened to personal freedom? #StopHillary • Don't allow this republic to become a monarchy. #StopHillary #StopBill • Democrats say that sanctuary cities increase public safety? #StopHillary	
Abortion	En route to Parnell Square for the rally for life!! #everylifematters	NEGATIVE	NEGATIVE	NEUTRAL	• RT [USER] It's almost time! #rally4life #everylifematters • Please help support The Rally for Life 2015, add a #Twibbon now! • RT [USER] Heartbreaking Video Shows the Sign Language for Abortion #prolife #tcot	
Feminism	Rather be an "ugly" feminist then be these sad people that throws hat on people that believes in equality!	POSITIVE	POSITIVE	NEGATIVE	• She should have all the right to an abortion! It's her body, her baby! It's her choice, not yours! #prochoice • #MarriageEquality #LoveWins • [Why does media keep saying "the 21 years old man blablabla". Who cares if he's 21. He's ****ing a racist terrorist! A TERRORIST. #charleston]	
Abortion	Today is a great because we are alive.	NEUTRAL	NEUTRAL	NEGATIVE	• [USER] we are so thrilled to have you speak at your banquet on May 1! Counting down the days... #prolife #lifeforall • Today, let's pray for life. #prolife #abortion #abortionhurts	

Figure 1: We look at some examples in the stance detection task where HuFT-Llama$_{LHLC}$ predicted the correct stance and Llama$_{LHLC}$ was incorrect. We highlight some of the selected text from the respective author's context that suggests having helped the HuFT model better understand language within the author's context.

We load the respective pre-trained adapted weights into the bigger HuLM models and fine-tune the same PEFT modules as in pre-training for each downstream task. We train all models for 5 epochs with a learning rate of 5e-5 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) with a batch size of 4.

For consistency, we use 512 tokens per instance with a batch size of 32 for non-HuLM baselines. We optimize training using cross-entropy loss for classification tasks and mean squared error loss for regression tasks.

5.3 Fine-tuning for Downstream Tasks

Here, we adopt the standard fine-tuning approach where the model is given one 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 asking the model to independently predict a person-level attribute for each document written by the author and averaging the predictions across all documents to arrive at a person-level prediction. We use this approach to compare Llama, Llama$_{LHLC}$, and HU-Llama over downstream performance.

We set up the training environment to replicate that of HuFT, and for consistency, use 512 tokens per instance with a batch size of 32. We optimize training using cross-entropy loss for classification tasks and mean squared error loss for regression tasks for both FT and HuFT.

5.4 Hardware

We use a pair of NVIDIA H100 80GB GPUs for continued pre-training of HU-Llama, HaRT$_{Llama}$, and Llama$_{LHLC}$. We run fine-tuning experiments on a single NVIDIA H100 80GB or an RTX A6000 48GB GPU.

6 Results and Discussion

6.1 Effect of HuFT on Larger LLMs

In general, we find HuFT for downstream tasks to perform better than fine-tuning without the author's context. HuFT-Llama$_{LHLC}$ performs better than Llama and Llama$_{LHLC}$ in four downstream tasks and is similar in the fifth task (see Table 4). Collectively processing an author's language helps across all person-level tasks. For document-level tasks, stance detection shows substantial gains owing to the personal nature of the task, whereas sentiment detection has no statistically significant difference in performance.

6.2 Role of Scale in HuLM Models

We find HuLM training and fine-tuning (HU-Llama) to benefit downstream performance, but simply HuFT to be similar in most tasks (see Table 5). We also find that scaling up the HuLM-based HaRT's recurrent architecture to a bigger Llama model—HaRT$_{Llama}$—does not show benefits over the non-recurrent HuLM-based HU-Llama. Interestingly, we find smaller-scale recurrent HuLM models (HaRT$_{GPT2\text{-}twt}$ and HaRT$_{GPT2\text{-}twt\text{-}fb}$) to perform better on the document-level tasks of stance and sentiment detection.

6.3 Role of Author's Context

Empirical Findings Regarding the Amount of Author's Context. We further assess the effect of the amount of author's context on the performances of smaller- and bigger-HuLM-based models. We find that not using the author's context when fine-tuning HU-Llama affects the performance on all downstream tasks, achieving similar results as Llama$_{LHLC}$ not-HuFT (see Table 6). This further supports the benefits of HuFT that we discuss in Section 6.1. Additionally, while an author's context helps HU-Llama's downstream performance, further increasing the context length when fine-tuning (see HU-Llama$_{8192}$ in Table 6) does not provide significant gains in results for the tasks of estimating age (in blogs) and occupation, where additional author's context was available.

Qualitative Analysis. We look at some examples from the stance detection and occupation classification tasks to find where the models benefit from processing language within the author's context as opposed to when text documents written by the same author are processed independently. We show some of the examples in Figure 1 and 2 suggesting Llama$_{LHLC}$ better understands language within the author's context, i.e., when we adopt HuFT for respective downstream tasks.

6.4 Comparison with Prompting

Prompting performs poorly (see Table 7) for person-level classification (i.e., occupation) and regression tasks (i.e., age estimation) where Llama instruction-tuned versions are prompted with historical language from an author to estimate the

Model	Stance(F1)			Occupation(F1)		
	Seed 42	Seed 1234	Seed 3	Seed 42	Seed 1234	Seed 3
Llama$_{LHLC}$	57.83	60.46	61.63	53.74	53.50	54.41
HU-Llama	67.39	67.54	67.49	57.69	57.76	58.06

Table 8: Role of random seeds: We run experiments with three seeds for Llama$_{LHLC}$ and HU-Llama across two downstream tasks, and find similar trends in results.

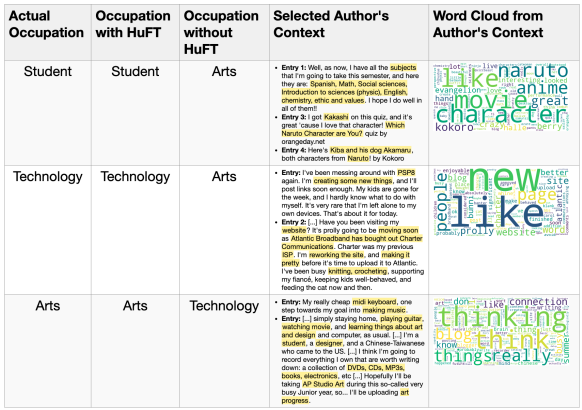

Figure 2: We look at some examples in the occupation classification task where HuFT-Llama$_{LHLC}$ predicted the correct label and Llama$_{LHLC}$ was incorrect. We highlight some of the selected text from the respective author's context that suggests having helped the HuFT model better understand language within the author's context.

person's job or age. This is potentially due to the lack of training the model to process language in the author's context, and being inadequate in performing tasks at the person-level.HuFT-Llama$_{LHLC}$ performs substantially better than prompting for the person-level tasks, further providing evidence for impressive improvements when fine-tuning within the author's context. We note here that we use the instruction-tuned version of Llama, which is known for its strong performance at document-level tasks. While not the main question under our study, we see marginal gains in using the author's context when prompting for stance detection.

6.5 Randomness and Hardware Variability

We test for the effects of random seeds on our model performances by running experiments with three seeds using Llama$_{LHLC}$ and HU-Llama and find similar result trends. We observe a degree of variability in the results depending on the type of GPU used, and we note it as an infrastructural limitation for our study.

7 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 user. 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. Pre-training with author context based language modeling objective did not yield additional benefits in the 8B models unlike with the smaller 110M parameter GPT-2 model. These results together demonstrate the utility of modeling the primary generators of language, the 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 LLMs. Thus, we note that assessing the full impact of HuLM pre-training in larger LLMs 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 whole other research question yet to be explored and beyond the scope of our study. Furthermore, our study's scope does not include prompt engineering or assessing the efficacy of prompting in various conditions with the author's context. We include basic prompting experiments for completeness of comparison only.

Ethical Considerations

The multi-level human-document-word architecture of HuLM enables large language models to in-

corporate 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.

References

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of WASSA 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525, Toronto, Canada. Association for Computational Linguistics.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, Dublin, Ireland. Association for Computational Linguistics.

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.

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, João Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024c. Findings of WASSA 2024 shared task on empathy and personality detection in interactions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 369–379, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins,

Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wen-

wen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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 Jimenez, Daniel King, Pengyu Liu, Steven Wu, Graham Neubig, Barnabás Póczos, Yonatan Bisk, and Shashank Srikant. 2023. Swe-bench: Can language models resolve real-world github issues? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

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.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

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.

Karan Singhal, Shekoofeh Azizi, Tao Tu, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, C. J. Humeau, and et al. 2022. Large language models encode clinical knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024. 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.

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 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 111–119.

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 Prompting

We use llama3.1-8b-Instruct-hf model and the vLLM framework (Kwon et al., 2023) for our prompting experiments. For prompting, we use 2 different methods (with and without author context). We are using llama3.1-8b-Instruct-hf model for prompting. Prompt details are given in Table 9.

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	Identify the stance of the given target text towards {topic}. Select one of the three: In Favor, or Against, or Neutral. Here is a list of the previous messages written by the person in chronological order to learn more about the person: {messages} 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	Identify the sentiment of the given target text. Select one of the three: Positive, or Negative, or Neutral. Here is a list of the previous messages written by the person in chronological order to learn more about the person: {messages} Here is the target text: {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 9: Topic = [Hillary Clinton, atheism, feminism, legalization of abortion, climate change as a real concern], {messages} are separated by a line.