Comparing Pre-trained Human Language Models: Is it Better with Human Context as Groups, Individual Traits, or Both?

Nikita Soni[1], **Niranjan Balasubramanian**[1], **H. Andrew Schwartz**[1], and **Dirk Hovy**[2]
[1]Stony Brook University, [2]Bocconi University
{nisoni, niranjan, has}@cs.stonybrook.edu, dirk.hovy@unibocconi.it

Abstract

Pre-trained language models consider the context of neighboring words and documents but lack any author context of the human generating the text. However, language depends on the author's states, traits, social, situational, and environmental attributes, collectively referred to as human context (Soni et al., 2024). Human-centered natural language processing requires incorporating human context into language models. Currently, two methods exist: pre-training with 1) group-wise attributes (e.g., *over-45-year-olds*) or 2) individual traits. Group attributes are simple but coarse — not all 45-year-olds write the same way — while individual traits allow for more personalized representations, but require more complex modeling and data. It is unclear which approach benefits what tasks. We compare pre-training models with human context via 1) group attributes, 2) individual users, and 3) a combined approach on five user- and document-level tasks. Our results show that there is no best approach, but that human-centered language modeling holds avenues for different methods.

1 Introduction

Language is a fundamental form of human expression that varies between people. Pre-trained Language Models (PLMs) account for the textual context of neighboring words and documents but lack the human context of the author "generating" the language. However, language is highly dependent on the human context (Soni et al., 2024), i.e., an author's changing states (Fleeson, 2001; Mehl and Pennebaker, 2003), traits, social, situational, and environmental attributes. For example, a person's language differs when hiking (situation/environment) versus when feeling dejected (state) over a breakup (situation). It is essential to model the additional *human context* to better understand human language with PLMs (Soni et al., 2024). Two strands of human-centered Natural Language Processing (NLP) emerged to model the people behind the language. The first focuses on the *group context*, building on the sociolinguistic notion of specific socio-demographic attributes influencing the language of a particular group. These socio-demographic attributes include age, gender (Volkova et al., 2013; Hovy, 2015), location (Kulkarni et al., 2016; Garimella et al., 2017), personality (Schwartz et al., 2013; Lynn et al., 2017), and more. The second approach focuses on building personalized language models that target *individualistic contexts* (King and Cook, 2020; Delasalles et al., 2019), and latent attributes inferred from an individual's historical language (Matero et al., 2021; Soni et al., 2022) to better model the user.

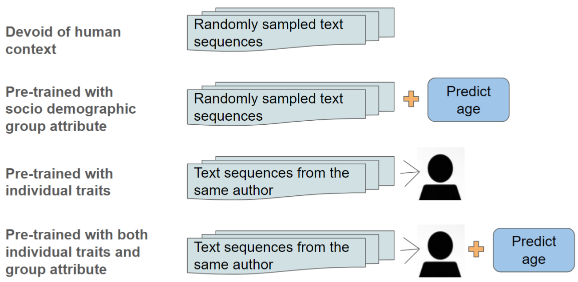

Figure 1: Pre-training a language model with no human context, with socio-demographic group attribute, with individual traits, and with both group and individual traits.

While these two strands have advanced human-centered NLP, we still do not understand their relative strengths, complementarity, and impact on different tasks (Soni et al., 2024). People are not defined by their group memberships alone (Orlikowski et al., 2023), and individual traits might not be generalizable enough across groups. Further, cross-cultural psychology research (Hofstede and Bond, 1984) notes the importance of both individualism and collectivism and Soni et al. suggest a flexible interplay of these concepts when including human context in PLMs. We might expect models

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 316–328
August 15, 2024 ©2024 Association for Computational Linguistics

informed by both group and individual features to perform better, but no data exists on this. In this work, we empirically evaluate these hypotheses and compare the effects of including different types of human context in PLMs (i.e., groups context (collectivism), individualistic aspects (individualism), and a combination of both) on specific tasks. We answer the following broader research questions:

***(RQ1)**: How can we incorporate group and individual human context into pre-training?*

***(RQ2)**: How does pre-training with different human contexts affect downstream performance for different tasks?*

Recent works trained PLMs with socio-demographic group context (Hung et al., 2023), and individual context (Soni et al., 2022). However, no empirical studies compare the impacts of different types of human contexts included in language modeling. We compare the downstream tasks' performance of models from these works with that of a new PLM trained with group and individual contexts. We test all systems on five downstream tasks from these works to ensure a variety of tasks and prediction properties at three levels: 1) user level, with historical language from authors (age estimation and personality assessment), 2) document-level, with historical language from some authors (stance detection), 3) document-level, without historical language from authors (topic detection and age category classification).

Note that because we focus on empirically comparing pre-training with different types of human context, we cannot compare to the larger language models like GPT4, which are not yet pre-trained/trainable with human context but are envisioned to become large human language models in the future (Soni et al., 2024). Recent studies have explored methodologies like user-adapters (Zhong et al., 2021) and user-centric prompting (Li et al., 2023) to include human context into the larger language models. In contrast, we focus specifically on comparing the impact of *pre-training* LMs with different human contexts, as Soni et al. (2024) argue that pre-training will allow for modeling a richer human context by explicitly handling the multi-level structure of documents connected to people, as opposed to specific and limited benefits of user-centric prompting and fine-tuning (Salemi et al., 2023; Choi et al., 2023).

PLMs pre-trained on individual *and* group features enhance performance on two user-level regression tasks that use multiple user documents: age estimation and personality assessment. In contrast, PLMs pre-trained on *individual* human context alone improve performance on document-level classification tasks like stance and topic detection. Our findings suggest user-level tasks focusing on individual people are best modeled as a mix of both group attributes and unique characteristics. However, document-level tasks that are more personal, like stance detection, are best addressed by modeling the individual context alone. Adding group attributes degrades performance.

By their very nature, models of this kind touch upon sensitive user information. For this reason, we adopt a responsible release strategy, making only the code for the comparisons publicly available, along with the exact splits of the TrustPilot and Stance datasets used[1]. We build on top of the publicly available code from Soni et al. (2022) and Hung et al. (2023). We acquired the models and data from the authors of the respective works in a secure manner. For more information about the models and data, see Sections 4 and 5. For a discussion of the ethical implications of the models and data, see the Ethical Considerations section.

Contributions. (1) We empirically compare three pre-training strategies for language models with human context: individual traits, group socio-demographic features, and a combination of both.
(2) We evaluate each pre-training strategy on five downstream tasks: two multi-document user-level regression (personality-openness evaluation and age estimation) and three single document-level classification tasks (stance detection, topic detection, and age category classification).
(3) We find that the two user-level regression tasks perform better with PLMs pre-trained with individual and group human contexts. Conversely, the three single document-level tasks perform better with PLMs pre-trained with individual context alone. Further, pre-training with group and individual contexts reduces performance for document-level tasks.

2 Related Work

Socio-demographic and latent human factors. Much work in human-centered NLP has focused on identifying and evaluating inclusion of human context in our models. Initial studies show benefits of grouping people by socio-demographic factors

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

like age or gender (Volkova et al., 2013; Hovy, 2015) and geographical region (Bamman et al., 2014; Garimella et al., 2017) to capture the variation in language usage and meaning among different groups, and improving text classification tasks like sentiment analysis. Additionally, adapting to socio-demographic user factors (Lynn et al., 2017), social networks (Huang et al., 2014; Radfar et al., 2020), and social media attributes (Bamman and Smith, 2015) have been effective to enhance the performance in tasks like sarcasm detection, and toxic language detection. Some studies go beyond modeling explicit groups, to learn individual representations latently Jaech and Ostendorf (2018); Delasalles et al. (2019) or via historical language Matero et al. (2021).

Pre-traing with human context. With respect to pre-trained LMs, recent studies have used adapter-based methodology (Li et al., 2021; Zhong et al., 2021) to include individual human contexts for downstream tasks. More recently, large language models have used user-centric prompting (Li et al., 2023) to include human context and evaluate on personalized and social tasks, resulting in mediocre performance (Salemi et al., 2023; Choi et al., 2023). However, few studies have explored including human context within the pre-training regime of LMs. Hung et al. (2023) generalize the task-specific EMPATH-BERT (Guda et al., 2021) to create a PLM injected with demographic group information using a dynamic multi-task learning setup. Additionally, Soni et al. (2022) pre-train a LM with individual human context derived from user's historical language. Our study aims at comparing the impacts of pre-training LMs with individual, or group, or combined individual plus group human context.

3 Integrating Human Context in PLMs

For our comparison, we use three systems representing the three paradigms of pre-training with human context (Figure 1). We want to tease apart the contributions of different types of human context: 1) grouping people, 2) modeling individual users, and 3) modeling both group and individual human contexts. As noted earlier, we focus on recent approaches for pre-training language models with additional human context.

Pre-training with group context. We build on a model from Hung et al. (2023) that explores demographic adaptation in transformer-based PLMs. It is a bidirectional auto-encoder-based PLM injecting demographic knowledge in a multi-task learning setup where they train for masked language modeling (MLM) and classify the gender or age of an author. They use the Trustpilot dataset [2] of multilingual reviews with demographic labels (Hovy, 2015), and evaluate on multiple text classification tasks, including demographic attribute classification, sentiment analysis, and topic detection. For our comparison study, we use the US-English subset of the Trustpilot data for two tasks: topic detection (TD) across two age categories, and age attribute classification (AC) (more details in section 5). Additionally, we use a monolingual BERT pre-trained with age specialization on the Blogs authorship corpus (Schler et al., 2006). This choice allows us to eliminate a domain influence (i.e., Trustpilot reviews), given that the other PLMs under comparison lack this specialization.

Pre-training with individual human context. Soni et al. (2022) introduced human language modeling (HuLM) in PLMs, which is regular language modeling given an additional context of the individual generating the language. This additional context is a dynamic vector derived from the authors' historical texts motivated by the idea of capturing the changing human states expressed in language. It also adds coherence to texts generated by the same author. They introduce Human-aware Recurrent Transformer (HaRT), an autoregressive PLM to evaluate the effect of individual human context on language modeling and multiple user-level and document-level downstream tasks. We build on HaRT and use two user-level tasks, age estimation and personality (openness) assessment, and on a document-level task, stance detection, for our comparisons study.

Pre-training with both group and individual human context. We train a PLM to integrate both individual and group human context by introducing a multi-task learning setup into HaRT that incorporates group features. This approach facilitates training a PLM with both group and individual context. We evaluate the model on two multi-document user-level regression tasks: age estimation and personality assessment, and three single document-level classification tasks: stance detection, topic detection, and age group classification. Importantly, the only difference in this multi-task learning setup compared to HaRT is the

[2] https://www.trustpilot.com/

inclusion of a demographic attribute prediction during pre-training, similar to how Hung et al. (2023) adapted traditional PLMs for group context.

4 Models

4.1 Pre-training with individual human context

HaRT. Soni et al. (2022) use a 12-layered autoregressive GPT-2 based architecture with a modified self-attention computation at layer 2. This modification to the query vector now includes the individual human context via a dynamic user-state vector.

$$Q_i^{IN} = W_q^T [H_i^{(IN-1)}; U_{i-1}]$$

where IN is the insert layer (layer 2), Q_i is the query vector under computation, H_i is the hidden states vector, and U_{i-1} is the user-state vector derived from the previous block of language seen from the user. All the text from a user is processed in the same forward pass with recurrent processing of blocks of fixed-length (1024) tokens chunked after temporally ordering the social media posts by created time. The user state is recurrently updated using the hidden states from layer 11 and computed as follows:

$$U_i = tanh(W_U U_{i-1} + W_H H^{(E)})$$

where, E is the extract layer (layer 11), U_i is the updated user-state vector, U_{i-1} is the user-state vector from the previous block, and H^E is the hidden states vector from layer 11. This formulation of updating the user-state vector extends the previous user-state vector information with the current language block's information.

HuLM Pre-training Task. HaRT is pre-trained for the human language modeling (HuLM) task defined as predicting the next token given the previous tokens while conditioning on previous user state $U_{1:t-1}$ (Soni et al., 2022) .

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

This is translated into a pre-training objective to maximize:

$$\prod_{a \in \text{Users}} \prod_{t=1}^{|\mathcal{B}_a|} \prod_{i=1}^{|B_t^{(a)}|} Pr(w_{t,i}|w_{t,1:i-1}, B_{1:t-1}^{(a)})$$

where, $w_{t,i}$ is the i^{th} token in the t^{th} block ($B_t^{(a)}$) for user a. The tokens from the previous blocks are represented using HaRT's recurrently updated user-state vector. Soni et al. use cross-entropy loss for the HuLM objective.

4.2 Pre-training with group human context

BERT$_{\text{DS}}$ and BERT$_{\text{age-MLM}}$. Hung et al. (2023) explore socio-demographic adapted BERT models to inject group human context into PLMs. We use two models: one specialized for age (demographic attribute) under the multi-task learning setup, and the other adapted to the age corpora through standard masked language modeling. We denote these as BERT$_{\text{DS}}$ and BERT$_{\text{age-MLM}}$, respectively.

Multi-Task Learning. Hung et al. (2023) train for both domain adaptation using the masked language modeling (L_{mlm}) loss and for classifying demographic category using the binary cross-entropy loss (L_{dem}). Both losses must be combined to simultaneously learn multiple objectives. To account for the *homoscedastic uncertainty* (Kendall et al., 2018) of both losses, they adopt a dynamic multi-task learning (MTL) objective for training with group human context. Homoscedastic uncertainty is a task-dependent weighting to derive a multi-task loss function that can optimally learn the weights and balance the impact of multiple loss functions and their different scales. The tasks are dynamically weighted using the variance of the task-specific loss (σ_t^2) over training instances of the task $t \in \{mlm, dem\}$:

$$\tilde{L_t} = \frac{1}{2\sigma_t^2} L_t + \log \sigma_t$$

Hung et al. minimize the sum of both the uncertainty adjusted losses: $\tilde{L}_{mlm} + \tilde{L}_{dem}$.

4.3 Pre-training with both individual and group human context

GRIT. We train HaRT under a multi-task learning setup for both the individual context — through the HuLM pre-training task (see Section 4.1) — and the group features — via a regression task to predict a (continuous) socio-demographic attribute of the author. We call the model as **GR**oup and **I**ndividual HaR**T** (GRIT). The model uses the user-state vectors (see Section 4.1) to predict the socio-demographic attribute of the author:

$$Pr(attribute|\overline{\mathbf{U}})$$

We chunk a user's language history into blocks and process them in a single forward pass. Each block of text from a user results in a user-state vector. We use the average of the user-state vectors from each non-padded block of texts from an author to compute their final user-state representation. This representation is layer-normed and linearly transformed before making a continuous-valued prediction for the specific attribute.

We pre-train one model for the continuous attribute age (GRIT$_{\text{age}}$) and one for the continuous attribute personality type openness (GRIT$_{\text{ope}}$). The models train on a regression loss for the attribute prediction regression tasks using mean squared error loss (L_{mse}), and a classification loss for the HuLM task using cross-entropy loss (L_{ce}). We must combine both losses to jointly learn the two objectives and account for the *homoscedastic uncertainty* (Kendall et al., 2018) of the losses. Since we combine a regression and a classification loss, we train the model to learn to balance the loss for a continuous and discrete output as derived in Kendall et al. (2018) and compute our joint objective as follows:

$$\frac{1}{\sigma_{ce}^2} L_{ce} + \frac{1}{2\sigma_{mse}^2} L_{mse} + \log \sigma_{ce} + \log \sigma_{mse}$$

where, σ_{ce}^2 and σ_{mse}^2 are the variances of the task-specific losses over the training instances of the respective tasks.

To add numerical stability, we adjust the loss calculation to use log of the variance:

$$\exp^{-\eta_{ce}} L_{ce} + \frac{1}{2}(\exp^{-\eta_{mse}} L_{mse} + \eta_{ce} + \eta_{mse})$$

where $\eta_x = \log \sigma_x^2$ for $x \in \{mse, ce\}$. We let σ_{ce} and σ_{mse} be learnable parameters for the model. In practice, we do not halve the η_{ce} term in the above equation since we found it to perform better with our multi-task learning experiments.

Pre-training Data and Training. We use the same Facebook posts dataset (Park et al., 2015) and training, validation, and test splits as those used by Soni et al. (2022). For both GRIT$_{\text{age}}$ and GRIT$_{\text{ope}}$, we use the demographic and personality scores, respectively, obtained from consenting Facebook users (Kosinski et al., 2013). This data is identical to that used by HaRT for the age estimation and personality assessment tasks. During training, we use a learning rate of 5e-5 in the multi-tasking training setup, employing the homoscedastic loss computation method described earlier. Following the experimental settings for HaRT, each training instance is capped to 4 blocks of 1024 tokens each. We use a train batch size of 1 per device and an evaluation batch size of 20 per device, trained over 2 GPUs for 8 epochs. Further details can be found in Appendix A.1.

4.4 Fine-Tuning

We utilize the results of fine-tuned BERT$_{\text{DS}}$ and BERT$_{\text{age-MLM}}$ from Hung et al. (2023), as well as fine-tuned HaRT models from Soni et al. (2022) where available. We fine-tune both GRIT models for all downstream tasks, and HaRT for 2 document-level tasks. Additionally, we use the Optuna framework (Akiba et al., 2019) for hyperparameter search, closely following the experimental settings in Soni et al. (2022). Details can be found in Appendix A.2.

4.5 Transfer Learning

We experiment with fine-tuning GRIT$_{\text{age}}$ in a multi-task learning setup for both the HuLM task and predicting personality (openness). Similarly, we fine-tune GRIT$_{\text{ope}}$ to predict age while also training for the HuLM task. We observe that this form of transfer learning yields the best performance for the user-level regression tasks (refer to Section 6.1).

5 Experiments

Our study's goal is to compare the downstream performance of models pre-trained with human contexts in three forms: socio-demographic group factors, individual traits, and combined. To this end, we evaluate performances of the models defined in Section 4 on two multi-document user-level regression tasks: predicting age and a personality score (openness), and on three single document-level classification tasks: stance detection, topic detection, and age classification. We also compare against GPT-2$_{\text{HLC}}$ from Soni et al. (2022) as a PLM adapted to the social media domain but devoid of human context. All experiments were run using Optuna trials (Akiba et al., 2019) to search for the best hyperparameters and reduce the effects of randomness. More details are provided in Appendix A.2

5.1 User Level Regression Tasks

We consider two user-level social scientific tasks: age estimation, and personality (openness) assessment, which require predicting continuous outcomes (real-valued age, or openness score) for a user given multiple documents written by them. We use the same data splits as used by Soni et al. (2022) for our comparison study.

Since $GRIT_{age}$ is pre-trained using age estimation as one of the tasks, we use directly evaluate it on the held-out test set. This allows for direct comparison with HaRT fine-tuned for the age estimation task. Furthermore, we can potentially attribute performance differences to the training with combined group and individual context, as $GRIT_{age}$ incorporates the group feature into HaRT's architecture. Similarly, $GRIT_{ope}$ is evaluated on the held-out test set for personality assessment. Moreover, we evaluate $GRIT_{age}$ and $GRIT_{ope}$ for the tasks of personality assessment and age estimation, respectively, using the transfer learning mechanism described in Section 4.5. We report and compare pearson r for age estimation and disattenuated pearson r for personality assessment.

5.2 Document-Level Classification Tasks

We compare different models for stance detection vs. topic detection and age classification tasks. These tasks classify a single input document (tweet message or a review) that a user writes into label categories. For stance detection, we also utilize the historical messages of a user where available, as in Soni et al. (2022). However, we do not have the user information or any user historical language available for the other two tasks, so we evaluate solely based on the single document input.

All models process the input document(s) and feed the layer-normed last non-padded token representation to the classification layer to classify the document into label categories. Only GRIT and HaRT incorporate user information and the historical language available for the stance detection task. However, GPT-2_{HLC}, and both $BERT_{DS}$ and $BERT_{age\text{-}MLM}$ lack this hierarchical structure and can only use the input document without access to historical data for making predictions. We compare the results from Soni et al. (2022) and Hung et al. (2023) wherever applicable and fine-tune all the parameters of the respective pre-trained models and the classification heads for other task-model combinations using the standard cross-entropy loss.

Stance Detection Given a single annotated tweet, this task predicts a user's stance as in favor of, against, or neutral towards one of the five targets: atheism, climate change as a real concern, feminism, Hillary Clinton, and legalization of abortion. We fine-tune $GRIT_{age}$ and $GRIT_{ope}$ for each target separately, and use the results from Soni et al. (2022) for GPT-2_{HLC} and HaRT. We report the average of weighted F1 scores[3] with three labels across all five targets. We use the train/dev/test split provided by Soni et al. (2022) over the SemEval 2016 dataset (Mohammad et al., 2016). HaRT and GRIT models maintain the temporal accuracy by using only the messages posted earlier than the labeled messages from the extended dataset (Lynn et al., 2019) as a user's historical language.

Topic Detection We use the US subset of the TrustPilot reviews dataset (Hovy, 2015) from two age groups: below 35 or above 45 [4]. Given a single review, the task is to predict the review topics from five categories: Flights, Online marketplace, Fitness & Nutrition, Electronics, and Hotels. To maintain consistency, we adopt the same train, development, and test set splits as Hung et al. (2023) to ensure a stratified demographically-conditioned label distribution. We fine-tune GPT-2_{HLC}, HaRT, $GRIT_{age}$, and $GRIT_{ope}$ using these data splits to predict the topic for a given review, and report macro-F1 scores[3]. We also compare to results from $BERT_{age\text{-}MLM}$ and $BERT_{DS}$ (Hung et al., 2023).

Demographic Attribute Classification We use the same subset of the TrustPilot dataset as for topic detection and the same train, development, and test splits from Hung et al. (2023). Given a single review, this task predicts the age group binary label (<35 years old or >45 years old). Age categories are equally represented in each set. We fine-tune GPT-2_{HLC}, HaRT, $GRIT_{age}$ and $GRIT_{ope}$ using the provided splits to predict if the review is written by someone below 35 years or above 45 years, and report macro-F1 scores[3]. We also compare to results from $BERT_{age\text{-}MLM}$ and $BERT_{DS}$ (Hung et al., 2023).

5.3 Human Language Modeling

To compare the effects of individual and group factors on language modeling performance, we evaluate on the test set from the pre-trained data splits. We report and compare perplexity scores from the pre-trained GPT-2 (GPT-2_{frozen}), GPT-2_{HLC}, HaRT, $GRIT_{age}$ and $GRIT_{ope}$ for the human language modeling task.

[3]We use this metric to maintain consistency with previous works under comparison (Soni et al., 2022; Hung et al., 2023).

[4]As suggested by Hovy (2015), this split of the age ranges results in roughly equally-sized data sets and is non-contiguous, avoiding fuzzy boundaries.

Model	Human Context	*Age* (r)	*OPE* (r_{dis})
GPT-2$_{HLC}$	None	0.839	0.521
HaRT	Individual	0.868	0.619
GRIT$_{age}$	Ind + Grp	**0.890**	**0.658**
GRIT$_{ope}$	Ind + Grp	0.884	0.643

Table 1: Pearson r for age, disattenuated Pearson r for openness. Pre-training with individual plus group context show benefits in estimating age and assessing personality (openness). Bold = best in column. We find no statistical difference between GRIT$_{age}$ and GRIT$_{ope}$ for the task of age estimation. All other results show statistical significance $p < 0.05$ using paired t-test.

6 Results and Discussion

We report results for all the tasks here, discussing their respective impacts from pre-training LMs with individual human context, group context, and both individual and group context.

6.1 Comparisons Study

User-Level Regression Tasks. Table 1 shows the results of the two user-level regression tasks. We find that GRIT models outperform others for both age estimation and personality assessment tasks. Additionally, upon comparing the transfer learning (Section 4.5) outcomes of GRIT$_{age}$ for openness and GRIT$_{ope}$ for age to those of the HaRT and GPT-2$_{HLC}$ models, we consistently observe superior performance with the GRIT models, further substantiating their efficacy.

Note that while GPT-2$_{HLC}$ is a PLM that is adapted to the social-media domain, it lacks human context. HaRT incorporates individual human context in pre-training, and GRIT extends this by integrating both group and individual human contexts in pre-training (Figure 1). As shown in Table 1, there are gains observed from GPT-2$_{HLC}$ (no human context) to HaRT (individual human context), and further to GRIT (individual + group human context). This suggests that pre-training PLMs with individual and group human context can benefit multi-document user-level regression tasks, such as those we considered. Importantly, the only difference between HaRT and GRIT models lies in the integration of the demographic attribute prediction (group context). Both models are pre-trained and evaluated on precisely the same data, allowing performance differences to be attributed to the additional group context combined with individualistic human context.

Document-Level Classification Tasks. Table 2 shows the results for the 3 document-level classification tasks: stance detection, topic detection (TD) for 2 age groups (<35 and >45), and demographic attribute (age) group classification (AC). We see that task fine-tuned HaRT (individual human context) models perform better on all tasks.

HaRT models inherently include an additional context of the individual user and do not treat all inputs as if written by the same user. The considered stance detection task primarily relates to personal opinions and preferences, rather than group-level ones, making HaRT well-suited for incorporating such personalization due to its pre-training with individual human context. While a group context may also influence a person's stance to some extent, empirical observations show that the combination of individual and group contexts negatively impacts performance. Additionally, models pre-trained with group context (BERT$_{DS}$) perform well in group-based tasks such as topic detection and age classification. However, models pre-trained on both individual and group human context (GRIT) do not appear to enhance results in group-based, and personal stance detection tasks resulting in slightly worse performance.

Further, it is important to note that the individual human context (HaRT) derived for some of the users using their historical tweets, where available, in the stance detection dataset provides a richer human context as we see greater gains in the performance of HaRT over GPT-2$_{HLC}$. Conversely, when historical language is not available for certain datasets (topic detection and attribute classification), HaRT does not perform worse than GPT-2$_{HLC}$ and may even achieve marginal gains due to the inherent human context in the model. However, we leave the evaluation of the impact of historical language on human context for future work.

Perplexity. We also compare the language modeling capability of the various models. Table 3 reports perplexity on the held-out test set. The frozen GPT-2 performs poorly compared to the social media domain adapted GPT-2$_{HLC}$, while HaRT model with individual human context perform the best. In contrast, GRIT models with both individual and group human context exhibit a slightly lower perplexity than HaRT. An individual's language is inherently personal, yet it can also be influenced by their group context to some extent, thereby affecting the perplexity results in language modeling

Model	Human Context	*Stance* ($F1_{wtd}$)	*TD (<35)* ($F1_{mac}$)	*TD (>45)* ($F1_{mac}$)	*AC* ($F1_{mac}$)
GPT-2$_{HLC}$	None	68.60	69.77	65.43	63.93
BERT$_{age\text{-}MLM}$	Group	-	68.40	64.60	61.90
BERT$_{DS}$	Group	-	69.30	65.00	64.10
HaRT	Individual	**71.10**	**69.84**	**65.65**	**64.33**
GRIT$_{age}$	Ind + Grp	70.82	69.21	64.52	62.56
GRIT$_{ope}$	Ind + Grp	70.07	66.53	64.84	61.18

Table 2: Weighted F1 for stance detection, macro-F1 for topic detection (TD), and age classification (AC) on TrustPilot reviews. Pre-training with individual context appear to benefit all tasks. **Bold** = best in column; McNemar's test comparing classifiers does not show statistical significance between the best performing model (HaRT) and the best baseline with no individual context (GPT-2$_{HLC}$).

Model	Human Context	Test (*ppl*)
GPT-2$_{frozen}$	None	114.82
GPT-2$_{HLC}$	None	36.39
HaRT	Individual	**28.24**
GRIT$_{age}$	Ind + Grp	31.77
GRIT$_{ope}$	Ind + Grp	30.32

Table 3: Comparing perplexity on language modeling for models trained with individual and group contexts.

Age bucket	#Users	*HaRT (Ind)*	*GRIT$_{age}$ (Ind+Grp)*	*GRIT$_{ope}$ (Ind+Grp)*
<18	1113	0.223	**0.394**	0.393
18-21	1387	0.230	**0.278**	0.276
21-30	1557	0.512	**0.531**	0.519
30-45	695	0.485	**0.530**	0.520
45+	248	0.106	**0.205**	0.180

Table 4: Pearson r for age over five age buckets using different types of human contexts for error analysis. Bold indicates best in row. We find no statistical difference between GRIT$_{age}$ and GRIT$_{ope}$ for buckets 21-30 and 30-45. All other results show statistical significance $p < 0.05$ using paired t-test.

tasks. However, GRIT models pre-trained with both individual and group context yield slightly worse perplexity measures. Additionally, we observe similar trends in perplexity gains from GPT-2$_{HLC}$ (no human context) to HaRT (individual context) or GRIT (individual plus group context) as also demonstrated in Soni et al. (2022).

6.2 Error Analysis and Disparity

We conduct an error analysis based on a socio-demographic group attribute (age groups), specifically focusing on age and openness prediction tasks.

Task\Model	*HaRT (Ind)*	*GRIT$_{age}$ (Ind+Grp)*	*GRIT$_{ope}$ (Ind+Grp)*
Age (r)	0.215	**0.181**	0.185
OPE (r_{dis})	0.075	0.090	**0.072**

Table 5: Mean error disparity for age estimation and openness personality assessment over five age buckets. Bold indicates best in column (lower is better).

We measure the performance of GRIT and HaRT in terms of error disparity (Shah et al., 2020) — a systematic difference in error based on demographics as exemplified by the "Wall Street Journal Effect" (Hovy and Søgaard, 2015). We analyze both the prediction outcomes and error disparity in age and openness prediction for both models: HaRT, which considers individual context, and GRIT, which incorporates both individual and group context.

First, we split the task test dataset into different buckets based on the age groups (specifically, <18, 18-21, 21-30, 30-45, and >45 years old) of the users in the test set, and then we compare the performance of our models across these buckets. Results from Table 4 indicate that pre-training with individual and group contexts together performs better for estimating age across all the age groups, which implies it makes fewer errors as a function of the socio-demographic attribute age. We see similar trends for assessing openness personality (see Appendix Tables 6 and 8), suggesting that the group attribute prediction may act as a regularizer for models pre-trained with both individual and group contexts, thus aiding the models to make fewer errors across all age buckets.

To further confirm, we compute the mean error disparity (MED) as the sum of the differences in the performance metric (Pearson correlation for age, and disattenuated Pearson correlation for openness) across each pair of age buckets, which is then averaged by the number of pairs (Shah et al., 2020). A lower averaged sum of differences implies fewer errors as a function of the age groups. Lower MED scores for models pre-trained with individual and group context in Table 5 support our previous error analysis.

7 Conclusion

NLP benefits from modeling latent human context, such as socio-demographic group features or individual traits. A recent development has been to incorporate this additional human context into the pre-training regimen of LMs. However, humans exhibit varying degrees of group and individual characteristics. Understanding the impacts of pre-training with different types of human context will advance the integration of human context into our base LLMs (**?**). To assess the impacts, we compare three types of PLMs pre-trained with socio-demographic group attributes, individual human contexts, and combined group and individual traits, across five user- and document-level tasks. Our findings indicate that pre-training with both individual *and* group human context improves the two user-level regression tasks: age and personality prediction. Pre-training with individual human context enhances the performance of the three single-document classification tasks, including stance and topic detection. Interestingly, inclusion of both individual and group attributes results in reduced performance on the text classification tasks. Meanwhile, pre-training solely on group context aids in group-based document classification tasks, albeit suboptimally. These results represent a promising step towards modeling human context and offer valuable insights for the NLP community to investigate additional strategies for improving models with task-dependent human context during pre-training.

Limitations

The purpose of our study is to compare the impacts of modeling socio-demographic group attributes and modeling individual user traits, and we use relevant models to represent each of the approaches. There are likely to be other ways to model these approaches and the models we use are only one of the ways. Additionally, these models in themselves have limitations like the blocks mechanism to process all the text from author induces compute requirements resulting in a capping of the number of blocks used for training. While it is also unclear how many blocks are sufficient to capture the human context, and if it is helpful to use the earliest language or the most recently used language in the capped number of blocks.
Secondly, some of the datasets (TrustPilot) used do not have appropriate user identification or historical language to create an individual human context.
Lastly, as noted earlier, models and data that touch upon sensitive user information require an extremely responsible usage and limit researchers to make them publicly available.

Ethical Considerations

Models that incorporate socio-demographic information need to be considered with special scrutiny. On the one hand, they have the potential to produce fairer and more inclusive results, because they can account for human language variation. On the other hand, they risk revealing identifying or sensitive information, which can lead to profiling and stereotyping. These may present opportunities for unintended malicious exploitations. For example, models that improve demographic groups prediction or psychological assessments could be used for targeting content for individuals without their awareness or consent. Such models may also risk release of private information of the research participant if trained on private data unchecked for exposing identifying information. For this reason, we take a conservative release strategy. While we support open research and reproducibility, data and privacy protection take precedence. Thus, we will only be releasing the code for our comparison study and the data that does not contain sensitive information i.e., stance detection datasets and TrustPilot datasets for topic detection and attribute classification. This is also in accordance with the DUA we have received from the authors of the papers/models that we employ in our work.
Our comparison study aims to guide and further speed the growing body of human-centered AI research. The models under comparison aim to enable applicability in the interdisciplinary studies of the human condition leading to helpful tools for psychological health. However, at this point these models are not intended for use in practice and should be evaluated for failures. All user-level tasks presented here were reviewed and approved or exempted by an academic institutional review board (IRB). Our studies are limited to US-English

due to comparability reasons. However, similar effects are likely to hold for other languages, and should be evaluated in future work.

Acknowledgments

This research is 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, and a grant from the CDC/NIOSH (U01 OH012476). 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

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed Representations of Geographically Situated Language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland. Association for Computational Linguistics.

David Bamman and Noah Smith. 2015. Contextualized Sarcasm Detection on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):574–577. Number: 1.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Edouard Delasalles, Sylvain Lamprier, and Ludovic Denoyer. 2019. Learning Dynamic Author Representations with Temporal Language Models. *2019 IEEE International Conference on Data Mining (ICDM)*, pages 120–129. ArXiv: 1909.04985.

William Fleeson. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295, Copenhagen, Denmark. Association for Computational Linguistics.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3072–3079, Online. Association for Computational Linguistics.

Geert Hofstede and Michael H Bond. 1984. Hofstede's culture dimensions: An independent validation using rokeach's value survey. *Journal of cross-cultural psychology*, 15(4):417–433.

Dirk Hovy. 2015. Demographic Factors Improve Classification Performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy and Anders Søgaard. 2015. Tagging Performance Correlates with Author Age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.

Yu-Yang Huang, Rui Yan, Tsung-Ting Kuo, and Shou-De Lin. 2014. Enriching cold start personalized language model using social network information. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 611–617, Baltimore, Maryland. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavaš. 2023. Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1565–1580, Dubrovnik, Croatia. Association for Computational Linguistics.

Aaron Jaech and Mari Ostendorf. 2018. Personalized language model for query auto-completion. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 700–705, Melbourne, Australia. Association for Computational Linguistics.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Milton King and Paul Cook. 2020. Evaluating approaches to personalizing language models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2461–2469, Marseille, France. European Language Resources Association.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):615–618. Number: 1.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2023. Automatic prompt rewriting for personalized text generation. *arXiv preprint arXiv:2310.00152*.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human Centered NLP with User-Factor Adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthias R Mehl and James W Pennebaker. 2003. The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, 84(4):857.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934–952.

Bahar Radfar, Karthik Shivaram, and Aron Culotta. 2020. Characterizing Variation in Toxic Language by Social Context. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:959–963.

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.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE*, 8(9):e73791. Publisher: Public Library of Science.

Deven Santosh Shah, H Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264.

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. 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.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA. Association for Computational Linguistics.

Wanjun Zhong, Duyu Tang, Jiahai Wang, Jian Yin, and Nan Duan. 2021. UserAdapter: Few-Shot User Learning in Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1484–1488, Online. Association for Computational Linguistics.

A Appendix

A.1 Pre-training GRIT

Pre-training data. We use a subset of the pre-training data for HaRT, consisting of the demographics and personality information. This subset contains the Facebook posts from Park et al. (2015) as used by Soni et al.. Our dataset is consistent with the inclusion criteria for HaRT to ensure moderate language history for each user: we include English posts from users with at least 50 total posts and at least 1000 words. This dataset consists of just over 63,000 unique users, which we split into a training dataset consisting of messages from 56,930 users, a development dataset that consists of messages from 1836 users that were not part of the training set, and a test set of messages from a separate set of 4438 users that are neither in training nor the development set. To evaluate the human attribute prediction in GRIT_{ope}, we use a subset of the test set consisting of messages from 1745 users to accommodate for questionnaire reliability. We use the Facebook posts for the HuLM task and the demographic and personality scores of consenting Facebook users (Kosinski et al., 2013) for the human attribute prediction task.

Training. We use HaRT's pre-trained weights as the base weights for GRIT and randomly initialize the newly introduced weights for human attribute prediction. GRIT is trained on our pre-training dataset using the 5e-5 learning rate after experimenting with a few learning rates, including that used for HaRT's pre-training. Following HaRT, and due to computing limitations, each training instance is capped to 8 blocks of 1024 tokens each,

Age bucket	#Users	*HaRT (Ind)*	*GRIT$_{age}$ (Ind+Grp)*	*GRIT$_{ope}$ (Ind+Grp)*
<18	503	0.627	**0.644**	0.618
18-21	560	0.557	**0.608**	0.592
21-30	563	0.715	**0.741**	0.738
30-45	249	0.594	**0.669**	0.667
45+	68	0.567	0.546	**0.599**

Table 6: Disattenuated pearson r for openness over five age buckets using different types of human contexts for error analysis. Bold indicates best in row.

Age bucket	#Users	*HaRT (Ind)*	*GRIT$_{age}$ (Ind+Grp)*	*GRIT$_{ope}$ (Ind+Grp)*
<18	1113	4.07	**2.52**	2.82
18-21	1387	6.52	4.00	**3.89**
21-30	1557	17.82	**12.64**	13.11
30-45	695	48.59	**39.79**	40.43
45+	248	**114.92**	121.66	134.72

Table 7: Mean squared error for age over five age buckets using different types of human contexts for error analysis. Bold indicates best in row (lower error is better).

Age bucket	#Users	*HaRT (Ind)*	*GRIT$_{age}$ (Ind+Grp)*	*GRIT$_{ope}$ (Ind+Grp)*
<18	503	0.423	**0.410**	0.429
18-21	560	0.496	**0.487**	0.506
21-30	563	0.429	**0.380**	0.381
30-45	249	0.578	**0.489**	0.489
45+	68	0.584	0.501	**0.467**

Table 8: Mean squared error for openness over five age buckets using different types of human contexts for error analysis. Bold indicates best in row (lower error is better).

with train batch size as 1 per device and evaluation batch size as 20 per device, trained over 2 GPUs for eight epochs. We explored multiple joint losses before resorting to the homoscedastic loss computation. Since HaRT caps to 4 train blocks for user-level downstream tasks, we also pre-train $\mathrm{GRIT}_{\mathrm{age}}$ and $\mathrm{GRIT}_{\mathrm{ope}}$ with four training blocks.

A.2 Experimental Settings

We closely follow the experimental settings from Soni et al. (2022) and similarly use Optuna framework (Akiba et al., 2019) for hyperparameter search. We search for learning rates between 5e-6 and 5e-4, and between 1e-7 and 1e-5 for different tasks. We will make our best found hyperparameter values publicly available with our code and results in the github repository. All experiments are run on NVIDIA RTX A6000 GPUs of 48GB. Pre-training takes approx 14 hours for 1 epoch and fine-tuning takes approx 1-4 hours depending on the task.