Comparing Human-Centered Language Modeling: Is it Better to Model Groups, Individual Traits, or Both?

Anonymous ACL submission

Abstract

Natural language processing has made progress in incorporating human context into its models, but whether it is more effective to use group-wise attributes (e.g., *over-45-year-olds*) or model individuals remains open. Group attributes are technically easier but coarse: not all 45-year-olds write the same way. In contrast, modeling individuals captures the complexity of each person's identity. It allows for a more personalized representation, but we may have to model an infinite number of users and require data that may be impossible to get. We compare modeling human context via group attributes, individual users, and combined approaches. Combining group and individual features significantly benefits *user*-level regression tasks like age estimation or personality assessment from a user's documents. Modeling individual users significantly improves the performance of single *document*-level classification tasks like stance and topic detection. We also find that individual-user modeling does well even without user's historical data.

1 Introduction

Language varies between people. Two strands of human-centered NLP work have modeled the humans 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 attributes include socio-demographics like 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 strand focuses on building personalized language models (PLMs) 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.

While these two strands have advanced human-centered NLP, their relative strengths, complementarity, and impact over different tasks are poorly understood. People are not defined by their group membership alone (Orlikowski et al., 2023), but individual traits might not generalize. This paper compares the two approaches and their combination in the same framework and tasks in pre-trained large language models. We use the monolingual socio-demographically adapted model from Hung et al. (2023) and the HaRT model from Soni et al. (2022) for the first two types. We use a multi-task learning setup to create GRIT, a PLM based on HaRT, trained with both individual *and* group human context in two variants: GRIT_{age} is adapted to the authors' age, and GRIT_{ope} to their inferred personality trait (openness). We test all systems on five user- and document-level tasks. Note that because we focus on conceptually comparing group and individual traits, we cannot compare to GPT4, which does not support stratification to either attribute.

PLMs trained on individuals *and* groups enhance user-level regression tasks like age estimation and personality assessment from user's multiple documents. Such user-level tasks focus on individual person, and our findings show that these people are best modeled as a combination of their groups and individual traits conforming with the notion that a person is a mix of their group attributes and unique characteristics. Document-level categorization tasks, like stance detection, are more personal. A PLM taught within an individual human context alone improves our considered document-level classification tasks.

By their very nature, models of this kind touch upon sensitive user information. For this reason, we take a responsible release strategy, making only the code for the comparisons publicly available and the exact splits of the TrustPilot and Stance datasets used. We build on top of the publicly available code for HaRT and Hung et al. (2023).

We acquired the model and data in a secure manner from the authors of Soni et al. (2022) and TrustPilot data splits from the authors of Hung et al. (2023). For more information about the model and data, see Sections 3 and 4. For a discussion of the ethical implications of the model and data, see Section 7.

Contributions. Our contributions are: (1) We provide an analysis with a comparison of modeling individual users, group socio-demographic features, and both group and individual traits in PLMs. (2) We evaluate the three modeling stratEgoies on five downstream tasks: two user-level (personality-openness evaluation and age estimation) and three document-level classification tasks (stance detection, topic detection, and age category prediction). (3) We find that user-level regression tasks like estimating age and assessing personality from user's multiple documents perform better with mixed individual and group human context (GRIT) and document-level tasks like stance detection and topic detection perform better with individual context alone (HaRT).

2 Integrating Human Context in PLMs

For our comparative study, we use three systems representing the paradigms of human context modeling to tease apart the contributions of 1) grouping people, 2) modeling individual users, and 3) modeling both group and individual human contexts.

Training with group context. We build on Hung et al. (2023)'s work to explore demographic adaptation in transformer-based PLMs. They use bidirectional auto-encoder-based PLMs to inject demographic knowledge in a multi-task learning setup where they also train masked language modeling (MLM) and classify the gender or age of an author. They use the multilingual reviews dataset with demographic labels from Trustpilot[1] (Hovy, 2015). They evaluate multiple text classification tasks, including demographic attribute classification, sentiment analysis, and topic detection. In our study, we use the US-English subset of the Trustpilot data for topic detection (TD) under the age categories and for age attribute classification (AC) (more details in section 4). We use the results from Hung et al. (2023) with the monolingual PLM model BERT on a US-based English dataset with out-of-domain demographic (age) specialization for our comparison study. Out-of-domain data is the Blogs authorship corpus (Schler et al., 2006), and in-domain means Trustpilot corpus. To be consistent and fair in comparing with other human context training paradigms, we choose the monolingual model and eliminate the confounds from domain specialization.

Training with individual human context. Soni et al. (2022) introduced human language modeling (HuLM), i.e., training a regular language modeling task but including a dynamic individual human context vector derived from the authors' texts. This vector captures the human states in which the text is generated to induce coherence between different texts generated by the same author. Soni et al. also view this vector as representing the text-derived human factors. They introduce a Human-aware Recurrent Transformer (HaRT), a unidirectional autoregressive PLM that trains for the HuLM task. They evaluate the effect of individual human context on language modeling and multiple user-level and document-level downstream tasks. We use the results from HaRT on the user-level tasks, age estimation and personality (openness) assessment, and on a document-level task, stance detection, for our comparisons study.

Training with both group and individual human context. We train a PLM that can integrate the author's individual and group human context knowledge when training for language modeling. We extend Soni et al.'s HaRT model by training for HuLM and predicting a human group attribute in a dynamic muti-task learning setup as used by Hung et al.. We want to induce the individual human context through the author's language and inject the group context by predicting a group attribute of the author. Predicting group attributes during training can also be viewed as a regularizer for the model, as it constrains the possible output. We discuss the model in detail in section 3. We compare two user-level tasks and a document-level task, as in Soni et al. (2022), and on topic detection and age prediction tasks, as in Hung et al. (2023).

3 Models

This section describes the models we compare to represent group and individual human contexts.

3.1 Modeling individual human context

HaRT. Soni et al. (2022) introduced HaRT to incorporate individual human context into PLMs. They use a 12-layered autoregressive GPT-2 based architecture with a modified self-attention compu-

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

tation 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.

3.2 Modeling 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 the names BERT$_{\text{DS}}$ and BERT$_{\text{age-MLM}}$ to denote their demographic (age) specialization using the multi-task learning setup and demographic adaptation with masked language modeling respectively.

Multi-Task Learning. Hung et al. (2023) train for domain adaptation using the masked language modeling (L_{mlm}) loss and for classifying demographic category using the binary cross-entropy loss (L_{dem}). To account for the *homoscedastic uncertainty* (Kendall et al., 2018) of losses, they adopt a dynamic MTL objective for training with group human context. Kendall et al. interpret homoscedastic uncertainty as task-dependent weighting and derive a multi-task loss function that can optimally learn the weights to balance the impact of multiple loss functions. This approach accounts for the different scales across multiple losses.

$$\tilde{L}_t = \frac{1}{2\sigma_{mse}^2} L_t + \log \sigma_t$$

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

3.3 Modeling both individual and group human context

GRIT. GRIT incorporates both individual and group human contexts using a multi-task learning setup with HaRT that also predicts a socio-demographic attribute of the author in each forward pass using the average of user-state vectors from each non-padded block of the user's temporally ordered text.

Multi-Task Learning. GRIT is pre-trained for the HuLM task and a (continuous) socio-demographic attribute prediction regression task in a multi-task learning setup. The PLM uses the user-state vectors to predict the socio-demographic attribute of the user.

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

Because of compute limitations, 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 train one model for the attribute age (GRIT$_{\text{age}}$) and one for the attribute personality type openness (GRIT$_{\text{ope}}$), respectively. The models train on a regression loss for the attribute prediction regression tasks using mean squared error (L_{mse}), and a classification loss for the HuLM task using cross-entropy loss (L_{ce}). To account for the *homoscedastic uncertainty* (Kendall et al., 2018) of losses, we adopt a dynamic MTL objective as Hung et al. (2023). We use the joint 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} and σ_{mse} are the variances of the task-specific losses.

In practice, we use log variance for numerical stability and use the adjusted loss calculation as follows:

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

where $\eta_x = \log \sigma_x^2$. We let σ_{ce} and σ_{mse} be learnable parameters for the model. We do not halve the log-term of the cross-entropy loss since we found it to perform better with our multi-task learning experiments.

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$_{\text{ope}}$, 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, 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 GRIT$_{\text{age}}$ and GRIT$_{\text{ope}}$ with four training blocks.

3.4 Transfer Learning

To assess the efficiency of GRIT to transfer learning from predicting one group human attribute to another, we experiment with continuing pre-training for each group attribute. To this end, we pre-train GRIT$_{\text{age}}$ capped to 4 training blocks and use this pre-trained model to continue MTL with the HuLM task and predict personality (openness). We do the same for GRIT$_{\text{ope}}$ and continued MTL with predicting age.

3.5 Fine-Tuning

We fine-tune GRIT and HaRT for downstream document-level tasks. Each downstream task has a separate fine-tuned model that is initialized with the respective model's pre-trained parameters and trained using the respective downstream task labels and an appropriate loss function. We also use the historical language of a user where available for any of the downstream tasks. We use the last predicted token's representation to predict the label in document-level classification tasks. We experimented with fine-tuning GRIT for user-level regression tasks in multiple ways, including 1) similar to HaRT, by using the averaged user-state vectors from GRIT, 2) same as previous but fine-tuning only the history module, attribute prediction module, and the downstream task head, 3) freezing all the parameters of GRIT and fine-tuning the human attribute prediction module alone. However, we found continued training as described in section 3.4 to perform best.

We used the Optuna framework (Akiba et al., 2019) for hyperparameter search, closely following the experimental settings in Soni et al..

4 Experiments

We compare the performance of PLMs adapted to socio-demographic group factors, individual human contexts, and both individual and group contexts. We use the socio-demographic adapted

BERT models: $\text{BERT}_{\text{age-MLM}}$ and BERT_{DS} from Hung et al. (2023), HaRT (Soni et al., 2022), and GRIT models, respectively. We use $\text{GPT-2}_{\text{HLC}}$ from Soni et al. (2022) as a PLM adapted to the social media domain but devoid of human context. We evaluate performances on two user-level regression tasks: predicting age and a personality score (openness), and on three document-level tasks: stance detection, topic detection, and age classification.

4.1 User Level Regression Tasks

We compare GRIT, HaRT, and $\text{GPT-2}_{\text{HLC}}$ on age estimation and personality (openness) assessment. These tasks require continuous outcomes derived from multiple documents written by a user. We use a subset of the data from consenting Facebook users who shared their demographic and personality scores (Kosinski et al., 2012; Park et al., 2015) along with their Facebook posts. This data is essentially the same as GRIT's pre-training data. However, the test set is from Park et al. (2015) on which HaRT and $\text{GPT-2}_{\text{HLC}}$ are evaluated.

GRIT is pre-trained on a multi-task learning setup, including predicting a continuous socio-demographic group attribute. GRIT_{age} is trained to predict age, and GRIT_{ope} is trained to predict openness. We use the pre-trained GRIT_{age} and GRIT_{ope} directly to evaluate on the test sets for age estimation and personality assessment, respectively. We further evaluate these models on the test sets for personality assessment and age estimation after continuing training for these tasks, as described in section 3.4.

We use the results from Soni et al. (2022) for HaRT and $\text{GPT-2}_{\text{HLC}}$ which are directly comparable to GRIT models trained on the same data splits and metrics. Soni et al. fine-tuned the recurrence module of the pre-trained HaRT model for the tasks of age estimation and personality assessment using the average of user-states from non-padded blocks of texts from an author, resulting in two fine-tuned models. Similarly, they fine-tune the last two layers of the pre-trained $\text{GPT-2}_{\text{HLC}}$ model for these tasks. Since $\text{GPT-2}_{\text{HLC}}$ can not handle all text from a user in one pass, they average the predictions across all user messages corresponding to the same label for each message.

Age Estimation The training and development datasets are identical to the pre-training data for GRIT. Age was self-reported and limited to users under 65. We compare the performance of the models on a test set of 5000 users from Park et al. (2015) and report Pearson correlation (r).

Personality Assessment As for age estimation, the training and development datasets are the same as the pre-training data for GRIT. We compare the performance of the models to predict openness (one's tendency to be open to new ideas) on a test set of 1943 users and report disattenuated Pearson correlation (r_{dis}) metric to account for questionnaire reliability as in Soni et al. (2022).

4.2 Document-Level Classification Tasks

We compare different models for stance detection vs. topic detection and age classification. These tasks classify a single input document (tweet message or a review) a user writes into label categories. For stance detection, we also use the historical messages of a user where available, as in Soni et al. (2022). We do not have the user information or any user historical language available for the other two tasks, so we evaluate 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 (where available). The other two models can only use the input document without a hierarchical structure to make the predictions. We compare with 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 the models under comparison for each target separately. We report average of weighted F1 scores with three labels across all five targets. We use Soni et al. (2022)'s train/dev/test split over 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. We compare the results of fine-tuned $\text{GPT-2}_{\text{HLC}}$, HaRT (Soni et al., 2022), and fine-tuned

GRIT$_{age}$ and GRIT$_{ope}$.

Topic Detection We use the US subset of the TrustPilot reviews dataset (Hovy, 2015) from two age groups: below 35 or above 45 [2]. Given a single review, this task predicts the review topics from five categories: Flights, Online marketplace, Fitness & Nutrition, Electronics, and Hotels. We use the same train, development, and test set splits as Hung et al. (2023) to eliminate any skew in the demographically-conditioned label distribution. We report and compare macro-F1 scores from BERT$_{age\text{-}MLM}$ and BERT$_{DS}$ (Hung et al., 2023) with fine-tuned GPT-2$_{HLC}$, HaRT, GRIT$_{age}$ and GRIT$_{ope}$.

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. Age categories are equally represented in each set. We report and compare macro-F1 scores from BERT$_{age\text{-}MLM}$ and BERT$_{DS}$ (Hung et al., 2023) with fine-tuned GPT-2$_{HLC}$, HaRT, GRIT$_{age}$ and GRIT$_{ope}$.

4.3 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.

5 Results and Discussion

We report results for all the tasks here, discussing their respective impacts on adapting PLMs to individual human context, group context, and both individual and group context.

5.1 Comparisons Study

User-Level Tasks. Table 1 shows the two user-level task results. For computational reasons, both HaRT and GRIT use just 4 blocks of training data for these tasks. We find that the pre-trained GRIT models for each task perform better than the fine-tuned HaRT model, i.e., pre-trained GRIT$_{age}$ better estimates age, and GRIT$_{ope}$ better assesses personality. In addition, GRIT performs better on the other user-level task using transfer learning and continued training than by fine-tuning it on the averaged

[2]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. Bold = best result per column, $\dagger = p < .05$ (permtuation test w.r.t HaRT).

user states. Comparing the transfer learning results of GRIT$_{age}$ for openness and GRIT$_{ope}$ for age to the fine-tuned HaRT and GPT-2$_{HLC}$ models, we find that training PLMs with individual and group human context benefits such multi-document user-level regression tasks. Consequently, we may view the group attribute prediction as a regularizer for the model.

Document-Level Tasks. Table 2 shows the results for stance detection. Both GRIT and HaRT models were fine-tuned with historical language for stance. The PLM trained with individual human context (HaRT) better detects user stance.

Table 2 also compares performances of the models on topic detection (TD) and demographic attribute classification (AC). We use GRIT models pre-trained with 8 training blocks of user texts. Both HaRT and GRIT models have a notion of the user and treat each input as written by a different user. This aspect may aid their performances even for tasks where historical language is unavailable. Even though this information is lacking for TD and AC, fine-tuned HaRT models perform better than the rest. The additional group-attribute pre-training of GRIT models may be introducing noise for document-level tasks, since the results are close to that of fine-tuned HaRT models, yet slightly lower. We can draw parallels between the performance enhancements from GPT-2$_{HLC}$ to HaRT, i.e., a PLM adapted to an out-of-domain corpus (social media) to a PLM trained with individual human context using the same corpus, and between the performance gains from BERT$_{age\text{-}MLM}$ to BERT$_{DS}$, i.e., a PLM adapted to an out-of-domain corpus (blogs) to a PLM trained with group context using the same corpus. Note that Hung et al. (2023) do not report results from BERT$_{age\text{-}MLM}$ and BERT$_{DS}$ out-of-domain demographically specialized models. Our results indicate that single-document annotated classification tasks may benefit simply by

Model	Human Context	*Stance* (*$F1_{wtd}$*)	*TD (<35)* (*$F1_{mac}$*)	*TD (>45)* (*$F1_{mac}$*)	*AC* (*$F1_{mac}$*)
GPT-2$_{HLC}$	None	68.6	**69.8**	65.4	63.9
BERT$_{age\text{-}MLM}$	Group	-	68.4	64.6	61.9*
BERT$_{DS}$	Group	-	69.3	65.0	64.1*
HaRT	Individual	**71.1**†	**69.8**†	**65.6**	**64.3**†
GRIT$_{age}$	Ind + Grp	70.8	69.2	64.5	62.7
GRIT$_{ope}$	Ind + Grp	70.1	66.5	64.8	61.2

Table 2: Weighted F1 for stance detection, macro-F1 for topic detection (TD), and age classification (AC) on TrustPilot reviews. GPT-2$_{HLC}$ and HaRT from Soni et al. (2022), BERT$_{age\text{-}MLM}$ and BERT$_{DS}$ from Hung et al. (2023). * = results from in-domain specialized models. **Bold** = best in column; † = statistically significant $p < .05$ via permutation test w.r.t 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	503	0.223	**0.394**	0.393
18-21	560	0.230	**0.278**	0.276
21-30	563	0.512	**0.531**	0.519
30-45	249	0.485	**0.530**	0.520
45+	68	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.

training PLMs with individual human context.

Perplexity. We also compare the language modeling capability of the various models. Table 3 reports perplexity on the test set of 4438 users from the pre-training data. Multi-task learning is known to impact individual task performance, so we expect a slight dip in perplexity. The results align with our hypothesis and the trend shown in Soni et al. (2022). The frozen GPT-2 performs poorly compared to the social media domain adapted GPT-2$_{HLC}$, HaRT models perform best while GRIT models result in a slightly lower perplexity than HaRT.

5.2 Error Analysis and Disparity

We perform a set of error analysis by comparing performance metrics of HaRT and GRIT models (pre-trained with 4 training blocks) for the user-level regression tasks of age and openness prediction across different groups based on a demographic factor. The different groups are created by sampling the test set into the following age buckets: below 18, 18-21, 21-30, 30-45, and above 45. Table 4 shows GRIT$_{age}$ performs better for the task of estimating age across all age groups i.e., exhibits lesser error. We also see lesser errors in GRIT models for the openness assessment tasks (Appendix Table 6) as well as conforming results on both tasks when comparing using the MSE metric (Appendix Tables 7 and 8).

Additionally, we use the error analysis results to compare the error disparity (Shah et al., 2020) in GRIT models and HaRT. Error disparity can be exemplified by the "Wall Street Journal Effect" – a systematic difference in error as a function of demographics (Hovy and Søgaard, 2015). It can be calculated as the difference in the computed metric across different groups based on a demographic factor (Shah et al., 2020). We compute the mean error disparity (MED) as the sum of the differences in the metric (Pearson correlation for age, and disattenuated Pearson correlation for openness) computed for each group averaged by the number of difference pairs.

Table 8 reports the MED for each model-task pair for HaRT, GRIT$_{age}$, and GRIT$_{ope}$ models, and age estimation and openness assessment tasks. We find GRIT models to demonstrate lower mean error disparity for each metric i.e., making less error as a function of the age groups.

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

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

6 Related Work

People use language to communicate and convey meaning more than mere words. Much work in human-centered NLP has focused on identifying and evaluting including human context in our models. Initial studies experimented with grouping people by socio-demographic factors 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. These works improved sentiment analysis, polarity classification, and topic detection. Other researchers explored human factors like social networks (Huang et al., 2014; Radfar et al., 2020), occupation (Preoţiuc-Pietro et al., 2015), personality (Schwartz et al., 2013; Lynn et al., 2017), and social media attributes (Bamman and Smith, 2015; Lynn et al., 2019) to improve toxic language detection, sarcasm detection, or stance detection.

Some studies go beyond explicit groups and learning individual representations latently or via historical language. Jaech and Ostendorf (2018) learned latent user embeddings for search query completion. Delasalles et al. (2019) conditioned a language model on a recurrently updated latent author representation. Welch et al. (2020) motivate personalized word embeddings by jointly learning a latent representation for each user and generic word representations. Hofmann et al. (2021) combined learned latent representations of the social space with time to produce dynamic contextualized word embeddings. (King and Cook, 2020) created personalized models using the authors' historical text. Lynn et al. (2020) attend to user's past messages to better predict user personality.

Research on adapting pre-trained language models to socio-demographic factors has been minimal. Guda et al. (2021) propose EMPATH-BERT, a demographically-aware model to predict empathy and distress better. Lauscher et al. (2022) probe PLMs to understand if their representations encode socio-demographic information. Hung et al. (2023) generalize the task-specific EMPATH-BERT to create a PLM injected with demographic group information using a dynamic multi-task learning setup. We adapt their mono-lingual BERT-based model to age with out-of-domain data for our comparison.

Several studies (Li et al., 2021; Mireshghallah et al., 2022; Zhong et al., 2021) have explored adapting pre-trained Transformer-based language models to individual human contexts for downstream tasks. Li et al. (2021) find benefits in adding user IDs to the input to generate explanations for recommender systems. Zhong et al. (2021) learned a latent user-specific vector prepended to the input embeddings to classify sentiment better, similar to Mireshghallah et al. (2022), who use static user text identifiers instead. Matero et al. (2021) perform masked language modeling on users' past messages with message-level attention, producing efficient document representations for stance detection. Soni et al. (2022) propose human language modeling, where language is modeled conditioned on a dynamic user state derived from temporally ordered past user utterances. We use their state-of-the-art model HaRT in our comparison and as a base for our pre-trained model with individual and group human context.

7 Conclusion

NLP benefits from modeling group traits like sociodemographic factors and individual users in terms of latent human context. However, humans exhibit varying degrees of group and individual characteristics. Through a comparative study of five user- and document-level tasks, we uncover how using individual traits and group characteristics in PLMs optimizes user-level regression tasks like age estimation and openness assessment. Meanwhile, individual human context training alone appears to bolster single-document annotated classification tasks like stance and topic detection. Despite our progress, our research reveals there are still considerable strides to be made in modeling human factors in language models. Our findings provide valuable insight into including human context in pre-trained language models to suit specific applications.

Limitations

The purpose of our study is to compare the impacts of modeling sociodemographic 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 sociodemographic 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.

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.

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.

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.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Dynamic contextualized word embeddings. 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 6970–6984.

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.

Michal Kosinski, David Stillwell, Pushmeet Kohli, Yoram Bachrach, and Thore Graepel. 2012. Personality and Website Choice.

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.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020. Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316, Online. 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, 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.

Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim, and Dimitrios Dimitriadis. 2022. Useridentifier: Implicit user representations for simple and effective personalized sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3449–3456.

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.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. 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 1754–1764, Beijing, China. Association for Computational Linguistics.

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.

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.

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.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Exploring the Value of Personalized Word Embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6856–6862, Barcelona, Spain (Online). International Committee on 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.

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).

A Appendix

A.1 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.

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).