Representations from LLMs for Human-Centered NLP Tasks

Anonymous ACL submission

Abstract

Like most of NLP, models for human-centered NLP tasks – tasks attempting to assess author-level information – predominantly use representations derived from hidden states of transformer LMs. However, what component of the LM is used for the representation varies widely. Moreover, there is a need for Human Language Models (HuLMs) that implicitly model the author and provide a user-level hidden state. Here, we systematically evaluate the different ways of representing a document and a user using different LM and HuLM architectures over tasks capturing both dynamically changing and averaged user-level attributes of *valence*, *arousal*, *empathy*, and *distress*. We find that representing documents as an average of the token hidden states from the second to the last layer performs the best generally. Further, while a user-level hidden state itself is rarely the best representation, we find its inclusion in the model strengthens token or document embeddings used to derive user-level representations.

1 Introduction

Human-centered NLP focuses on the humans generating language (Flek, 2020; Hovy and Yang, 2021; Soni et al., 2024a) Many human-centered NLP tasks focus on assessing human-attributes of a user based on their language (Lynn et al., 2017), utilizing a representation of the author, often within a specific time period or set of documents (Matero et al., 2022; Giorgi et al., 2024).

Like most sub-fields in NLP, human-centered NLP heavily relies on LMs but they do not directly contain a representation for the person. For example, no particular layer, token, or output is accepted as the best way to represent a person. More recently, *human language models* (HuLMs) — those that directly model the human-level of generation (Soni et al., 2024b) — have been proposed but still lack a evaluations for capturing human-level factors.

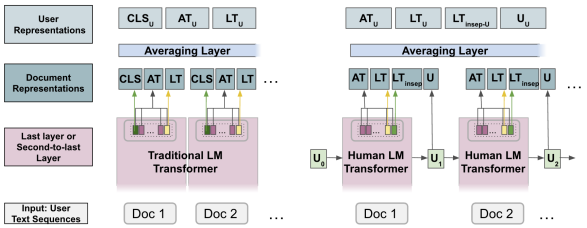

Figure 1: We evaluate representations derived from different hidden states in both traditional Transformer-based LMs and human language models (HuLMs). For document representations, we use the following methods: CLS, a special token prefixed to documents in auto-encoders; AT, averaged tokens in a document; LT, last token in a document for autoregressive models; LT_{insep}, last token as *insep* for HuLMs; and U, user state for HuLMs. We average the corresponding document representation types for all documents written by a user to derive their user representation.

Here, we systematically compare and evaluate different types of word-, document-, and human-level representations derived from standard LM architectures as well as from HuLM architectures (Figure 1) across multiple tasks spanning fundamental human attributes (the affective circumplex (Russell, 1980; Mohammad, 2016), empathy, and distress (Batson et al., 1987; Litvak et al., 2016)). Transformer-based LM (Vaswani et al., 2017) architectures provide us with many choices for representation: first token (i.e., CLS) (Devlin et al., 2019; Liu et al., 2019) in case of auto-encoders, the last token in case of autoregressive models (Radford et al., 2019), or the average of all the tokens in a document. Since traditional LMs do not have a notion of the user, prior works have used a hierarchical user representation by averaging all document embeddings written by the particular user (Nagao and Katsurai, 2024). While that is posthoc way to represent a user, HuLMs provide us with user states – a relatively more natural way for the same, by processing language in its multi-

level structure i.e., words are a part of a document, and documents are generated by people (Soni et al., 2022). We hypothesized the user states to help better assess the human-attributes.

Contributions. (1) We depict some of the different ways to represent a user, a user within a specific period of time (i.e., wave), and a specific document written by the user. (2) We systematically evaluate the multiple representation types over broadly 2 levels of analysis of user-level attributes: dynamically changing over time and somewhat stable on an average. (3) We draw 2 major findings: (a) averaging the token embeddings as the basis for document representations and a hierarchical user representations perform the best on an average, and (b) contrary to our hypothesis, user states themselves are not the best representations, but consistently strengthen user representations derived from token and document embeddings where a user-state was included.

2 Related Work

Human-centered NLP tasks are often grounded in the person and require user representations for downstream applications, such as assessing mental health (Ganesan et al., 2022) estimating demographic attributes (Hovy, 2015; Benton et al., 2016) , assessing personality (Schwartz et al., 2013), stance detection (Matero et al., 2021) and personalized recommendations (Qi et al., 2021; Wu et al., 2023). Commonly, a hierarchical approach has been used to average document representations across all text sequences from a particular user since the pre-LLMs era (He et al., 2019; Lynn et al., 2020). As Transformer-based LLMs become ubiquitous in human-centered NLP tasks, the hierarchical user representation approach continues to prevail (Lynn et al., 2020; Nagao and Katsurai, 2024). While this allows us to derive a user representation, it does not necessarily capture the *human context*, as traditional LLMs exhibit a form of ecological fallacy by either processing documents written by the same individual independently or, conversely, treating all language as if written by an average, universal person (Soni et al., 2024b).

Recently introduced Transformer-based Human Language Models (HuLMs) (Soni et al., 2022) offer a notion of the user by processing language in its multi-level structure, where words are part of a document, and documents are generated by people. This provides a relatively natural way of representing a user. Traditional LMs and HuLMs present a multitude of choices regarding which layer, token, or output to select for the best representation of a person and their document- and token-level representations. To this end, we perform a systematic study to identify the best representations.

3 Methods and Experiments

Architecture. We sought to use an architecture that can handle multiple criteria for testing representations: (1) multi-level structure spanning words, documents, and person-level, (2) last token and special token representations, (3) otherwise mirrors commonly used LLMs. To meet this criteria, for our primary results we built on Human-aware Recurrent Transformer (HaRT) that was pre-trained on the human language modeling task (Soni et al., 2022). HaRT was originally trained by concatenating temporally ordered text sequences from a user, separated by a special *insep* token, into blocks and each block was chunked at 1024 tokens (we call it $\text{HaRT}_{\text{frozen}}$). The user state vector updates after processing a block of tokens from that user. We use the language data (essays) from the DS4UD data (refer Section 4) and further train the publicly available pre-trained HaRT for the human language modeling task (and call it $\text{HaRT}_{\text{concat}}$). We also train another version that limits one block to consist of one document (and call it $\text{HaRT}_{\text{ODPB}}$). This enables us to get an updated user state after each text sequence written by a user at a given point in time.

Models and Types of Representations. Document and user representations as vectors are commonly used for predicting outcomes against a document, a user, or a specific time-period for a user (i.e., a wave). In this study, we will evaluate different types of representations from 4 models, of which 2 are autoregressive: 1) HaRT, 2) GPT-2 small, and 2 are auto-encoders: 3) BERT, 4) RoBERTa.

HaRT enables us to use 4 types of **document representations**: a) Last token (LT), b) Last token as *insep* (LT_{insep}), c) Average of token embeddings (AT), d) User states (U) (Figure 1). Autoregressive GPT-2 can provide the first 3 of these 4 types, while auto-encoders (BERT, RoBERTa) normally use the CLS token embeddings (CLS) or the AT to represent a document. We employ the hierarchical averaging structure over the different types of document embeddings across all documents from a user, or in a specific time-period (i.e., wave) to retrieve

Method	Dynamic States								User-Level				
	Document					Wave							
	Val	Aro	Emp	Dis	Avg	Val	Aro	Avg	Val	Aro	Emp	Dis	Avg
GPT-2													
LT	0.53	0.25	0.71	0.58	0.52	0.64	0.28	0.46	0.63	-0.14	0.37	0.42	0.32
AT	0.60	0.34	0.69	0.54	0.54*	0.77	**0.33**	**0.55***	0.75	0.18	0.52	**0.68**	0.53*
BERT													
CLS	0.58	0.30	0.70	0.66	0.56	0.74	0.28	0.51	0.71	0.09	0.45	0.42	0.42
AT	0.61	0.35	0.72	0.59	0.57*	0.75	0.26	0.50	0.73	0.14	0.45	0.39	0.43
RoBERTa													
CLS	0.62	0.35	**0.77**	0.63	0.59	0.75	0.29	0.52	0.72	0.10	0.51	0.59	0.48
AT	0.63	**0.36**	0.72	0.63	0.58*	0.75	**0.33**	0.54**	0.75	0.16	0.59	0.53	0.51
$HaRT_{frozen}$													
LT	0.52	0.22	0.70	0.57	0.50	0.69	0.21	0.45	0.73	0.18	0.42	0.60	0.48
LT_{insep}	0.55	0.23	0.76	0.67	0.55	0.69	0.22	0.46	0.74	0.23	0.45	0.62	0.51
AT	0.64	0.33	0.75	0.68	0.60*	0.79	0.27	0.53*	0.73	**0.35**	**0.67**	0.66	**0.60***†
U	-	-	-	-	-	-	-	-	0.59	-0.01	0.46	0.49	0.38
$HaRT_{concat}$													
LT	0.53	0.22	0.72	0.62	0.52	0.70	0.16	0.43	0.74	0.18	0.42	**0.68**	0.50
LT_{insep}	0.57	0.26	0.76	0.56	0.54	0.75	0.30	0.53	**0.77**	0.15	0.61	0.55	0.52
AT	**0.65**	0.34	0.75	**0.70**	**0.61***†	**0.80**	0.27	0.54	**0.77**	0.34	0.62	0.65	0.59*
U	-	-	-	-	-	-	-	-	0.65	0.06	0.59	0.49	0.45
$HaRT_{ODPB}$													
LT	0.57	0.27	0.68	0.56	0.44	0.72	0.17	0.52	0.72	0.24	0.41	0.55	0.48
LT_{insep}	0.64	0.34	0.74	0.66	0.59	0.79	0.27	0.53	0.72	0.09	0.40	0.50	0.43
AT	0.63	**0.36**	0.73	0.63	0.59*	**0.80**	0.28	0.54	0.75	0.12	0.58	0.51	0.49**
U	0.52	0.15	0.46	0.40	0.38	0.65	0.01	0.33	0.66	0.04	0.50	0.47	0.42

Table 1: We evaluate different types of representations from traditional LMs and human language models over tasks of predicting valence (Val), arousal (Aro), empathy (Emp), and Distress (Dis), both as dynamic states and averaged user-level attributes. Document-, wave-, and user- reperesentations based on averaged token embeddings (AT) generally outperform LT, CLS, LT_{insep}, and U across respective models. Additionally, token and document embeddings informed by the user state improve user-level representations, resulting in best results for averaged user-level outcomes. Results are reported in Pearson r, with bold indicating best in column and * indicating statistical significance with $p < 0.001$ and ** with $p < 0.05$ for each model's AT results against its second best result. † represents statistical significance ($p < 0.05$) across models for each of document-, wave-, and user-level analysis.

different types of **user and wave representations**, respectively.

Experiments. We use the different types of document representations to predict the dynamic outcomes – valence, arousal, empathy, and distress – against a person's respective text sequence. The valence and arousal outcomes are averaged for a person for a specific period of time (i.e. a wave) to get the outcome labels associated with a person for that wave. These wave-level outcomes are averaged across waves to get the respective labels associated with a person. The empathy and distress outcomes are averaged across all text sequences from a person to get a user-level label. We run ridge regression (with aplha values 1.e+01, 1.e+02, 1.e+00, 1.e+03, 1.e+04, 1.e+05, 1.e+06) over different types of representations derived from the four models (HaRT, GPT-2, BERT, and RoBERTa) to predict the respective outcomes. We average the results (Pearson r) across 10- straitified group based cross-validation folds such that no user overlaps between the train and test splits in any fold. The experiments took ~5 hours and were run on A6000 GPU utilizing approx ~24GB memory.

4 Datasets and Tasks

We select 2 datasets that allow for the multiple levels of analysis broadly categorized into (a) dynamic states (attributes that change with time and utterances), and (b) user-level (attributes on an average for a person). These datasets provide tasks of predicting fundamental human-attributes like affective circumplex grid (Russell, 1980), and empathy and distress (Batson et al., 1987).

Valence and Arousal. We use part of the dataset collected under the study DS4UD (Data Science for Unhealthy Drinking) that consists of essays written by US service industry workers on how they were feeling three times a day for 14 days, along with questionnaires where they select how they feel on the affective circumplex grid among many other attributes (Nilsson et al., 2024). This was done 6 times over a period of 3 years (2021 to early 2024). Each set of the 6 times is referred to as a "wave" and the user is asked to write essays and fill in questionnaires three times a day for 14 days, but required to do so only once a day (i.e., a users may have one to three essays per day). We use the self-reported outcomes derived from the affective

circumplex grid: *Valence* with labels ranging from 0 to 4 (highly negative affect to highly positive affect), and *Arousal* with labels ranging from 0 to 2 (low energy to high energy). We include users who participated in at least 2 waves and wrote at least 2 essays and each essay with at least 10 words, resulting in 120 users overall, 406 user instances across 6 waves (a user can have 2 to 6 instances, depending on the number of waves they participated in), and 10090 essays across all users. We compute outcomes labels across waves by averaging the labels corresponding to each essay for a user in a particular wave. Further, we average the labels corresponding to each wave for a user to get the user-level outcomes labels, following classical test theory (Rust and Golombok, 2014).

State Empathy Prediction. We use part of the dataset from the WASSA 2024 Shared Task on Empathy and Personality detection in Interactions, specifically a part of Track 3 (Giorgi et al., 2024). The task is to predict the empathic concern and personal distress scores (Batson et al., 1987) for each essay written by a crowd worker about the news article they read. Both empathy and distress scores have real valued labels between 1 and 7. We include train and dev data from WASSA 2024 and WASSA 2023 in our experiments – the years that have user-level data – resulting in 180 users and 1837 essays. We average the empathy and distress scores corresponding to each essay written by a particular user to get the user-level outcomes labels.

5 Results and Discussion

Averaged Token Embeddings (AT). Using the averaged token embeddings helps more often than the last token or CLS embeddings for all types of models considered: autoregresive, auto-encoders, and HuLMs. We performed initial evaluation by fetching hidden states from last layer and second to last layer, and comparing the different types of embeddings over the document, wave, and user-level tasks of predicting the human attributes (valence, arousal, empathy, and distress) (Appendix Table 2). We find that averaged token embeddings are best derived from the second to last layer from all models, whereas the last layer provides better embeddings for the other types – last tokens and CLS tokens. We report the results from the corresponding layers in Table 1 that shows for each model AT gives the best results on an average for all tasks.

Inclusion of Human Context. The user states by themselves are rarely the best embeddings but consistently the LT and AT-based embeddings were strong when a user-state was included. We see under the HuLMs (HaRT) in Table 1, representations from U to perform below par, however user-level representations derived from AT (and in some cases last tokens LT and LT_{insep}), that are processed in the context of the user states, boost the performance. While using AT is the current best choice, this also opens up avenues for future works to improve the natural representations of U and thus further improve the word's meaning embedded in tokens.

Last token as a special token (LT_{insep}). LT_{insep} is able to capture more information than last token. Interestingly, also under the HuLMs (HaRT) in table 1, we find representations derived from LT_{insep} embeddings to perform better than those from LT embeddings that suggests *insep* is able to capture the average of the tokens in a documents from a user.

6 Conclusion

We provided the first systematic evaluation of different types of token-, document-, and user-representations across multiple levels of analysis for tasks predicting both dynamically changing and averaged user-level attributes (valence, arousal, empathy, and distress). We used traditional LMs (BERT, RoBERTA, and GPT-2) as well as human language models (versions of HaRT) and represented documents using the CLS token, the last token, the last token as *insep*, U (user state from a *HuLM*), and the averaged tokens hidden states for respective models. We hierarchically average the corresponding document representations written by a particular user to derive their user representation. Generally, document-, wave- (user over a specific time period), and user- representations based on averaged token hidden states outperform CLS, LT, LT_{insep}, and U for respective models. Although the user states themselves do not prove to be a powerful representation for predicting user-level outcomes, they improve the averaged user-level performance when using user representations derived from user-state-informed tokens hidden states. Overall, our study and findings provide a standardized comparison and identify the best representations to select when applying LMs for human-centered NLP tasks.

Ethical Considerations

Models that incorporate the human context and/or use data involving private and sensitive information must adopt a responsible use and release strategy. While such models are essential for advancing research in human-centered NLP and understanding language in the context of the author, they may also pose unintended nefarious risks. For these reasons, the DS4UD data used in this study, which involves private user data, cannot be made publicly available due to ethical and privacy concerns. The affective grid circumplex tasks presented here were reviewed and approved or exempted by an academic institutional review board (IRB) as per the original dataset owners.

Limitations

The purpose of our study is to evaluate and standardize some of the commonly used Transformer-based approaches for representing users, their documents, and tokens in human-centered NLP tasks. We use some of the relevant and widely used models, exploring a few different types of representations. While our study focuses on specific models and methods, we encourage the NLP community to investigate lesser-explored approaches for modeling human context and deriving representations. We also acknowledge limitations inherent to these models and representations, such as context size and the extent of a user's historical language used as human context, which are beyond the scope for this study. Secondly, some users and essays in the WASSA dataset were duplicates or skewed. We removed the duplicates and retained the remaining data for evaluation. Finally, models and datasets involving sensitive user information, such as the DS4UD dataset, require careful and responsible usage. Due to privacy concerns, the authors of these datasets cannot make them publicly available.

References

C. Daniel Batson, Jim Fultz, and Patricia A. Schoenrade. 1987. Distress and empathy: two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55 1:19–39.

Adrian Benton, Raman Arora, and Mark Dredze. 2016. Learning multiview embeddings of Twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucie Flek. 2020. Returning the N to NLP: Towards Contextually Personalized Classification Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

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

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

Yun He, Jianling Wang, Wei Niu, and James Caverlee. 2019. A hierarchical self-attentive model for recommending user-generated item lists. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1481–1490, New York, NY, USA. Association for Computing Machinery.

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 Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and linguistic behavior and its correlation to trait empathy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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, 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, Albert Hung, and H. Andrew Schwartz. 2022. Evaluating contextual embeddings and their extraction layers for depression assessment. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 89–94, Dublin, Ireland. 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.

Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

Hiroyoshi Nagao and Marie Katsurai. 2024. Researcher representations based on aggregating embeddings of publication titles: A case study in a Japanese academic database. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 277–282, Bangkok, Thailand. Association for Computational Linguistics.

August Håkan Nilsson, Hansen Andrew Schwartz, Richard N Rosenthal, James R McKay, Huy Vu, Young-Min Cho, Syeda Mahwish, Adithya V Ganesan, and Lyle Ungar. 2024. Language-based ema assessments help understand problematic alcohol consumption. *PLoS ONE*.

Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. HieRec: Hierarchical user interest modeling for personalized news 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 5446–5456, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

James Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

John Rust and Susan Golombok. 2014. *Modern psychometrics: The science of psychological assessment*. Routledge.

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.

Nikita Soni, Lucie Flek, Ashish Sharma, Diyi Yang, Sara Hooker, and H Andrew Schwartz. 2024a. Proceedings of the 1st human-centered large language modeling workshop. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*.

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. 2024b. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8623–8638.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems*, 41(1):1–50.

Method	Dynamic States											
	Document								Wave			
	Val		Aro		Emp		Dis		Val		Aro	
	L	SL	L	SL	L	SL	L	SL	L	SL	L	SL
GPT-2												
LT	0.53	0.48	0.25	0.18	0.71	0.71	0.58	0.58	0.64	0.61	0.28	0.15
AT	0.60	0.60	0.34	0.34	0.70	0.69	0.58	0.54	0.76	0.77	0.29	0.33
BERT												
CLS	0.58	0.59	0.30	0.29	0.70	0.71	0.66	0.60	0.74	0.70	0.28	0.29
AT	0.61	0.61	0.35	0.35	0.71	0.72	0.59	0.59	0.77	0.75	0.29	0.26
RoBERTa												
CLS	0.62	0.59	0.35	0.30	0.77	0.69	0.63	0.42	0.75	0.71	0.29	0.30
AT	0.63	0.63	0.36	0.36	0.71	0.72	0.60	0.63	0.79	0.75	0.33	0.33
$\text{HaRT}_{\text{frozen}}$												
LT	0.52	0.51	0.22	0.20	0.70	0.70	0.57	0.56	0.69	0.68	0.21	0.16
LT_{insep}	0.55	0.56	0.23	0.21	0.76	0.74	0.67	0.52	0.69	0.69	0.22	0.20
AT	0.63	0.64	0.33	0.33	0.77	0.75	0.70	0.68	0.79	0.79	0.27	0.27
U	-		-		-		-		-		-	
$\text{HaRT}_{\text{concat}}$												
LT	0.53	0.52	0.22	0.18	0.72	0.70	0.62	0.60	0.70	0.70	0.16	0.13
LT_{insep}	0.57	0.59	0.26	0.26	0.76	0.46	0.56	0.48	0.75	0.74	0.30	0.28
AT	0.64	0.65	0.34	0.34	0.77	0.75	0.72	0.70	0.80	0.80	0.30	0.27
U	-		-		-		-		-		-	
$\text{HaRT}_{\text{ODPB}}$												
LT	0.57	0.53	0.27	0.21	0.68	0.63	0.56	0.52	0.72	0.69	0.17	0.13
LT_{insep}	0.64	0.65	0.34	0.33	0.74	0.73	0.66	0.65	0.79	0.80	0.27	0.25
AT	0.63	0.63	0.36	0.36	0.74	0.73	0.64	0.63	0.80	0.80	0.29	0.28
U	0.52		0.15		0.46		0.40		0.65		0.01	

Table 2: We evaluate the performance differences in hidden states from the last layer and second to last layer for all models. We find second to last layer to perform better for AT hidden states, and last layer to work generally well for the other types of representations across models.