REFINING TEXT INPUT FOR AUGMENTATIVE AND ALTERNATIVE COMMUNICATION (AAC) DEVICES: ANALYSING LANGUAGE MODEL LAYERS FOR OPTIMISATION

Hussein Yusufali, Roger K. Moore, and Stefan Goetze
{hsayusufali1, r.k.moore, s.goetze}@sheffield.ac.uk

Speech and Hearing Research Group, Dept. of Computer Science, University of Sheffield, UK

ABSTRACT

Communication impairments are prevalent among a significant proportion of individuals. Methods of Augmentative and Alternative Communication (AAC) can support people with speech disorders (PwSD) to some extent, but AAC users encounter substantial difficulties when engaging in open-domain social interactions, especially involving multiple participants. This is mainly due to the significant communication rate gap between typical speakers and AAC users. Large Language Models (LLM) offer a solution by providing predictions of the next words or sentences. This work analyses refining the prediction capabilities of Masked Language Models (MLM) for AAC users by performing layer-wise analysis specifically for word prediction on an AAC corpus. Experiments show that fine-tuning only specific low-performing LLM layers leads to better results than fine-tuning of the entire model. Fine-tuning of specific layers of a Robust Bidirectional Encoder Representations from Transformers (RoBERTa) model outperforms other tested models; for qualitative evaluation and informal prototype AAC device testing. Fine-tuning the word predictions in an AAC context results in approx. 20% increase in average communication rate (across different communication scenarios) to input speed of approx. 30 words per minute (WPM).

***Index Terms*—** Language Modelling, Augmentative and Alternative Communication (AAC), Text-to-Speech (TTS), Communication Rates

1. INTRODUCTION

It is estimated that over 70 million individuals globally have a particular type of communication impairment [1, 2], for instance, dysarthria, trauma to the vocal apparatus (such as cancer, laryngectomy, glossectomy), substantial impairments to disfluency, i.e. stuttering or stammering; hindering their abilities to communicate naturally. Consequently, individuals who face significant communication impairments can benefit from using AAC devices and strategies. The extent to which AAC devices are utilised and their perceived benefit for users are influenced by the severity of the individual's impairment [3, 4]. AAC users either require low-technology systems such as Pragmatic Organisation Dynamic Display (PODD) boards or communication charts or high-technology strategies such as Voice Output Communication Aids (VOCA) or text prediction devices [5]. Nonetheless, AAC devices still do not allow individuals to participate in multiparty social interactions in real-time successfully, primarily because present devices lead to extended pauses, produce inaccurate and unnatural predicted utterances, and lack fast, interactive and expressive capabilities needed to engage effectively with conversation partners [6, 7].

A significant inhibiting factor in the use of AAC devices and the successful participation by AAC users in interactions are low communication rates provided by current AAC devices. Communication rates in comparison to spoken conversational rates can be $5 - 15$ WPM, which is significantly lower than $120 - 140$ WPM for typical speakers [7, 8]. To alleviate the negative consequences of this large communication rate gap, text prediction techniques [9] can be utilised in AAC devices to circumvent the communication delays faced by users.

Utterance-based devices (UBDs), where the user selects from predefined phrases, bridge the communication rate gap; however, they lack user flexibility [10, 11]. Keyboard-based AAC devices circumvent this drawback and allow literate users to type; this, together with integrated Text-to-Speech (TTS) functionality, consequently enables the user to communicate freely. Nevertheless, the achievable typing rates are still not comparable to spoken conversational rates, with a maximum typing rate of approx. 45% WPM [12]. Recently, there have been several attempts to enhance the text-entry rates for spoken generative keyboard-based AAC devices, for instance, leveraging contextual information [13] or sentence abbreviation [14]. Nonetheless, these advances are still not comparable to natural multiparty interactions.

Statistical language modelling techniques have been prevalent in predictive text input for AAC devices and can potentially increase communication rates [15–17]. Predictions can either be based on character, word or phrase-based inputs. Fine-tuning language models such as Bidirectional Encoder Representations from Transformers (BERT) [18], RoBERTa [19] or Generative Pre-trained Transformer (GPT) [20] can enhance performance on specific downstream tasks. These foundational, transformer-based models have shown to be successful for a range of Natural Language Processing (NLP) tasks. However, they have mostly *black-box* character and are difficult to interpret. A further obstacle is their lack of reliability, transparency and predictability of their outputs [21, 22] especially for tasks like AAC use where they were not explicitly developed for. This is mainly due to resource-intensive demands, a lack of personalisation, and difficulties in ease of use. These language models' predictive capabilities often fail to convey the users' intentions or requirements accurately.

The main contribution of this paper is to optimise fine-tuning by layer-wise analysis of Large Language Models (LLM) specifically for AAC word prediction. Additionally, fine-tuned language models are evaluated for an AAC text input w.r.t layer-wise fine-tuning; by developing a prototype VOCA to test communication rates, across communication scenarios specifically designed for AAC usage.

In the following, the experimental methodology is described in Section 2, the outcomes and analysis are presented in Section 3, and conclusions, limitations and future scope is described in Section 4.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. This work was also funded in part by Apple Inc.

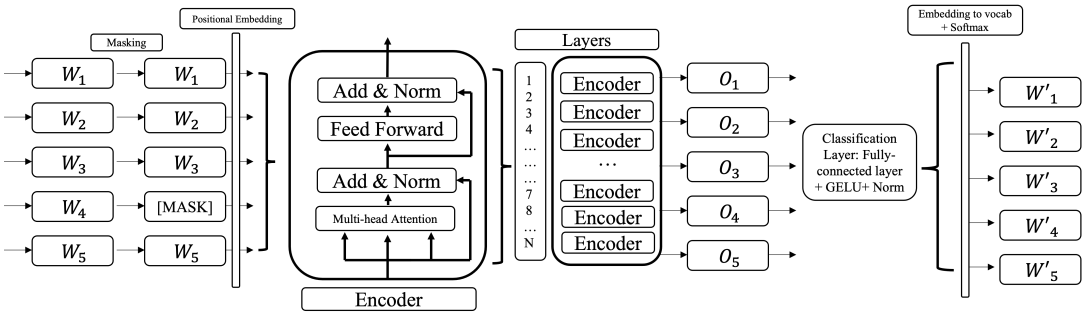

Fig. 1. Architecture of a LLM such as BERT or RoBERTa. Representing an individual encoder layer and its subsequent components, together with the input and output embeddings.

2. METHODOLOGY

2.1. Model Description

Figure 1 shows the MLM structure used in this work. The sequence of tokens W_i is the input of N transformer-based encoder layers (encompassing broad context). Typically 15% of the tokens are masked during self-supervised training for BERT [18] and RoBERTa [19] models. The encoder comprises multi-head self-attention and fully connected feed-forward layers. The attention mechanism maps queries and key-value pairs to generate an output. The overall model architecture compromises several encoder layers, where the initial input embedding representations are transformed and modified. Further to the N encoder layers, the prediction of the output words after a fully connected classification layer calculates the encoder outputs, then embedding the outputs to the vocabulary dimension (each output vector is multiplied by the input embedding matrix), and finally, a Softmax layer calculates the probability of the predicted words, represented by the $W^{'}$ in Figure 1.

These transformer-based architectures provide the ability to visualise how transformations of the tokens occur within the model, thus allowing visualisations as to which linguistic features are prominent and transformed within each encoder layer.

2.2. Corpora

AAC users are commonly subject to interactional asymmetry, where the user accommodates the conversational participants' communicative medium, typically speech. This is particularly prominent in social multiparty open-domain interactions, where the AAC user often requires long pauses or completely omits their conversational turns [7]. Currently, AAC systems are not yet fit for use in open-domain real-time interactions. Predictive systems, thus, have to be optimised specifically for open-domain conversations for AAC users. LLM are usually trained on extensive amounts of labelled data, however, such data is usually not available from communication of PwSD and communicative style and sentence structure of PwSD is not reflected. In this work, a recent, small AAC corpus[1] [23] is utilised to fine-tune the language models specifically for the downstream task of open-domain text prediction tailored to AAC users. The AAC corpus contains approx. 150k labelled utterances. It was found in previous work [24] that more generic corpora such as TV [25] or Reddit corpora [26] also lead to better text prediction results, however that the AAC corpus [23] achieves best AAC word prediction performance for PwSD.

[1] https://www.aactext.org/imagine/

The corpus was pre-processed by removing repeated letter words and abbreviations. Furthermore, the text was stemmed and punctuation removed, as retaining punctuation could negatively impact the usability due to increased time delays in scanning the generated predictions by the user. Additionally, any numerical predictions were filtered to prevent potential further time delays.

2.3. Layer-wise Analysis

Layer-wise analysis of the LLM involves isolating one specific encoder layer within the language model architecture whilst freezing all the other layers. A training loop is run using the AAC corpus training and test sets, and repeating this process for each layer. A batch size of 4 is used to mitigate memory consumption issues. The number of epochs is optimised depending on examining the training loss curves of the models. Training losses and an average word accuracy on the test set (for all batches per epoch) are computed during the training loop. During training, the remaining layers are *'frozen'*, i.e. their weights are not updated, allowing for a focused evaluation of the fine-tuned layer's performance.

2.4. Fine-Tuning

After investigating layer-wise metrics (c.f. Section 2.6), training losses, and accuracies of each layer of the models, a more well-founded and systematic approach to fine-tuning is formulated, specifically for open-domain word prediction for AAC devices. The layers with the lowest training losses and highest word prediction accuracies are frozen during fine-tuning; the *frozen* layers leverage the pre-trained knowledge of the model. The layers with low word accuracies and high training losses were fine-tuned on the downstream task and corpora. The low-performing layers are selected explicitly by examining the layer-wise metrics and determining which layers are primarily affecting the overall performance; a threshold accuracy value is chosen, and any layers performing below this threshold value are targeted for fine-tuning, following [27]. The lower layers do not capture global contextual features of the downstream task, so only the higher low-performing layers are selected to fine-tune. After the initial fine-tuning stage with freezing specific layers, fine-tuning is conducted by fine-tuning the complete overall models with lower learning rates. The learning rates are optimised by employing a learning rate scheduler, a step decay learning rate, where the learning rate starts high; once training converges, the learning rates adapt accordingly.

2.5. Fine-Tuning Details

Specific layer fine-tuning on the AAC corpora was performed utilising a standard sentence token batch size of 32, with 15 and 30 epochs for

the base and large models, respectively. The Adam Optimizer was used with an initial learning rate of $2 \cdot 10^{-5}$, which linearly decayed over the epochs, using a learning rate scheduler. Fine-tuning the models took approximately $6-12$ hours on $4xA100$ $80GB$ NVIDIA GPUs, with 4 CPUs per node and a RAM of $520GB$. A *Gaussian Error Linear Unit (GELU)* activation function was employed, with a dropout probability of 0.1 across all the layers. Memory consumption issues are of significance issue in the fine-tuning processes. To address this, $fp16$ (half-precision) was used. Balancing sequence length and memory is crucial for effective model training, as longer sequences capture a broader context but also result in increased memory consumption. Additionally, shorter sequence lengths were tested, and a token sequence length of 8 was used.

2.6. Objective Evaluation of Models

During and after both phases of fine-tuning, an objective evaluation of the fine-tuned language models is performed, with evaluation metrics of perplexity, F1-Score, BiLingual Evaluation Understudy (BLEU) [28] and word error rate (WER). BLEU is typically a metric for machine translation; however, it is adapted for word prediction by comparing N-gram overlaps between candidate and reference sentences. The WER metric is specifically chosen, as AAC users currently experience significant pauses in multiparty interactions due to unsatisfactory and inaccurate presented predictions, resulting in additional time delays, which negatively impact the user. To calculate layer-wise word accuracies, within the training loop, the masked sequences are iterated through in the test dataset; for each batch, the model compares the prediction to actual tokens. The accuracy per batch is computed as the ratio of the corrected predicted non-masked tokens to the total number of non-masked tokens. The final word accuracy per layer is computed as the average of the test batch accuracies. The accuracy metric evaluates how well the model performs at the word level and distinguishes the performance of each layer.

2.7. Prototype testing

Further to objective evaluation, qualitative informal evaluation and testing are undertaken by the primary author (who is also an AAC user) to determine input speed in close-to-realistic scenarios. These evaluations involve integrating the fine-tuned language models with TTS output. A Graphical User Interface (GUI) was developed and subsequently tested across four communication scenarios to determine achievable WPM by PwSD which is shown in Figure 2. The communication scenarios aimed to replicate realistic situations [24] which AAC users commonly encounter and also aimed to test the open-domain capabilities of the fine-tuned language models. The scenarios comprised of (i) a scripted coffee ordering, (ii) a half-prompted small talk at the workplace dialogue, (iii) an open-ended question and (iv) a picture description task [29].

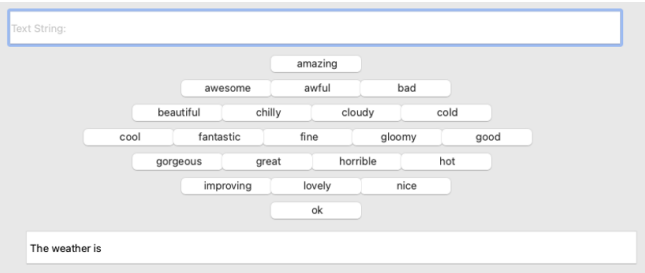

Fig. 2. Radial GUI with integrated incremental TTS as a prototype to test the efficacy of the fine-tuned language models, specifically for AAC use.

The communication rate gap is a factor of a combination of time delays [24]. To address and mitigate these, the prototype GUI shown in Figure 2 was developed and optimised to minimise unnecessary time delays. Several GUI layouts were designed to assess user interactions, cursor movements and input speed. A radial GUI was determined as the most efficient user interface [24]. Additionally, word-based TTS is utilised to decrease time delays further, and word-based subsequent predictions are presented to the user, contributing to smoother interactions experienced by the user and conversational participants. The prototype interface also allows the user to freely type if the presented predictions are not satisfactory.

3. RESULTS AND DISCUSSION

3.1. Layer-wise Analysis and Fine Tuning

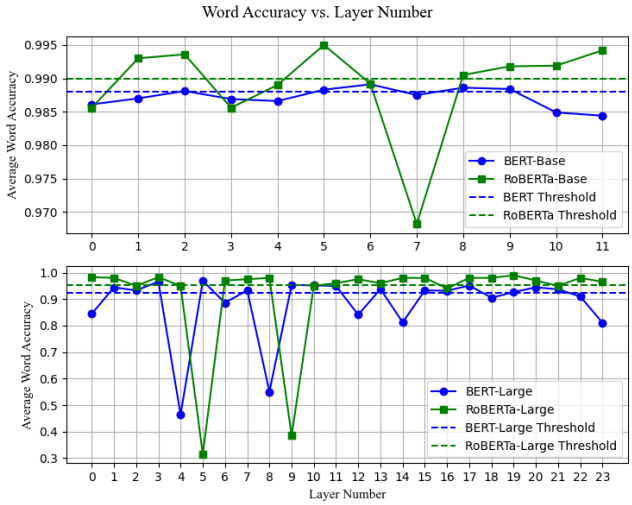

Fig. 3. Layer-wise analysis of both base uncased BERT and RoBERTa models, representing word prediction accuracies per layer. Dashed lines indicate empirically chosen thresholds to distinguish between better and lower-performing layers.

The results shown in Figure 3 confirm that not all encoder layers in these transformer language models perform consistently. The use of individual layers leads to different word accuracy values. However, due to the black-box nature of these models, the interpretability of why specific layers differ from others is challenging. The lower layers are usually assumed to capture syntactic and local contextual features, whereas the higher layers capture more abstract and global contextual information [30].

After fine-tuning specific layers, a comprehensive evaluation of the entire models was conducted. As shown in Table 1, both global (all layers) and specific layer fine-tuning improved all metrics, i.e. perplexity, F1-Score, BLEU and WER metrics across all the tested models. The best-performing model is the Base RoBERTa model, resulting in the best overall metrics across fine-tuned and frozen layers. The results also demonstrate how the RoBERTa base model leverages its pre-trained knowledge best against the AAC training corpora. The results distinctly convey how the training corpora of the LLM do not compare well to the communicative style of PwSD, with non-fine-tuned models performing worse across all metrics. Therefore, the optimisation of these models is beneficial. Table 1 also indicates that smaller Base models perform

Table 1. Objective evaluation of fine-tuned language models, particularly for AAC word prediction. GFT: Global Fine Tuning, LFT: Layer Specific Fine Tuning. Respective best performance indicated by bold-font.

Language Model	Accuracy Threshold	Fine-tuned Layers	Perplexity ↓	BLEU ↑	WER (%) ↓	F1-Score↑
BERT-Base	0.60	None	2.02	0.109	55.6	0.63
BERT-Large	0.63	None	2.17	0.098	58.9	0.61
BERT-Base GFT	0.71	All	1.56	0.212	42.8	0.72
BERT-Large GFT	0.77	All	1.42	0.299	46.4	0.76
BERT-Base LFT	0.988	[7,10,11]	1.102	0.397	26.92	0.87
BERT-Large LFT	0.945	[13,17,22,23]	1.347	0.233	31.56	0.84
RoBERTa-Base	0.62	None	1.98	0.130	52.3	0.66
RoBERTa-Large	0.65	None	2.13	0.102	56.7	0.69
RoBERTa-Base GFT	0.80	All	1.40	0.335	27.5	0.82
RoBERTa-Large GFT	0.73	All	1.39	0.301	33.6	0.79
RoBERTa-Base LFT	**0.990**	[6,7]	**1.075**	**0.401**	**23.32**	**0.91**
RoBERTa-Large LFT	0.952	[16,21]	1.146	0.356	25.98	0.88

better than the larger models in terms of all chosen metrics. This is particularly advantageous within the AAC field, as smaller models are more efficient to initialise and deploy on resource-restricted hardware. It should be noted that this outcome might have to be further analysed, as the difficulty primarily lies in users experiencing restricted capabilities in open-domain circumstances. Therefore, larger generative models could provide wider-ranging predictions. Hence, wider-ranging tests are needed, to experiment within a range of open-domain scenarios.

Complete model fine-tuning requires extensive training times, computational resources and training corpora for downstream task optimisation. Complete model fine-tuning on base models required greater than 12 hours on base models and approx. > 24 hours, on the large models. The results convey how fine-tuning specific layers is a cost-effective method for optimising models for a downstream task. A layer-specific fine-tuning process requires lower GPU training times and smaller specific corpora. Furthermore, the results also indicate how this could benefit the AAC field, which inherently needs more resources and data.

Further to the objective evaluation of the models, the results in Section 2.6 exemplified that specific layer fine-tuning outperformed all other models; therefore, the four best-performing models were integrated within the prototype AAC device.

3.2. Prototype Testing of Text-Prediction in an AAC System including TTS Output

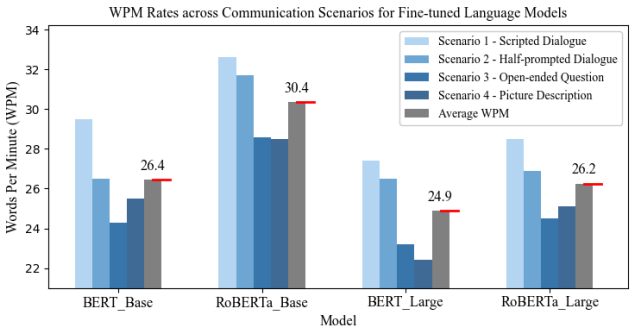

Fig. 4. The communication rates WPM, across four different communication scenarios and four different integrated language models when informally testing a prototype AAC system.

Figure 4 shows the performance of the text prediction in terms of achievable WPM when used in a VOCA for the four example dialogue tasks. The performance in Figure 4 is in line with objective performance metrics in Section 2.6. As expected, the open-ended questions and picture description tasks lead to average lower WPM since they are the more complicated scenarios. Again, the smaller models outperform the larger models. It should be noted that Figure 4 only shows results for an initial, informal test. Therefore, future work will focus on a more detailed, longitudinal testing of the language models integrated within AAC system.

Likewise, the communication scenarios described were designed to assess the open-domain nature of communication. Therefore, each dialogue was of a different prompted topic. Also, in this respect, the four currently chosen communication scenarios may only partially capture the wide-ranging, open-domain nature of AAC usage. Hence, more extensive studies might be necessary. During informal prototype testing, the evaluation indicated how word or phrased-based predictions overcome the restrictive nature of UBDs, as the prototype allows the user to type freely if the predictions do not accurately represent their intent.

The objective and qualitative evaluations conveyed that specific layer fine-tuning increased communication rates. On average, the communication rate between the four models was 27 WPM, with the base RoBERTa model achieving a maximum rate of 30.4 WPM, this does not surpass average typing speeds, however, does surpass other AAC text-input rates. Similarly, the evaluation also showed how crucial incremental TTS is for users; word-based TTS eliminates long pauses as experienced by whole utterance TTS. Testing also demonstrated how eliminating punctuation and numerical values had an overall positive effect in increasing the WPM rate, as less time was required to scan and review the generated predictions. However, further testing is needed to validate this outcome.

4. CONCLUSION

Communication impairments significantly inhibit PwSD in multiparty social interactions; this study demonstrates the potential of transformer-based language models in increasing the communication capabilities of text prediction in AAC system. The results show the importance of targeting specific encoder layers for optimisation instead of complete model fine-tuning, enhancing overall model performance and conserving computational resources through layer-wise analysis and fine-tuning. The results also highlighted that fine-tuning specific layers for a TTS downstream task can improve the communication rate of AAC systems. The outcomes, however, emphasise the necessity for larger-scale AAC corpora and longitudinal user testing to further refine text prediction for AAC models.

5. REFERENCES

[1] M. A. Morris, S. K. Meier, J. M. Griffin, M. E. Branda, and S. M. Phelan, "Prevalence and etiologies of adult communication disabilities in the United States: Results from the 2012 national health interview survey," *Disability and Health Journal*, 2016.

[2] J. Rennies, S. Goetze, and J.-E. Appell, "Personalized Acoustic Interfaces for Human-Computer Interaction," in *Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications*, M. Ziefle and C.Röcker, Eds., 2011.

[3] Z. C. Clarke, S. Judge, K. Fryer, S. Cunningham, J. Toogood, and M. S. Hawley, "A qualitative study exploring the effect of communicating with partially intelligible speech," *Augmentative and Alternative Communication*, vol. 39, no. 2, pp. 110–122, 2023.

[4] S. Goetze, N. Moritz, J.-E. Appell, M. Meis, C. Bartsch, and J. Bitzer, "Acoustic User Interfaces for Ambient Assisted Living Technologies," *Informatics for Health and Social Care, SI Ageing & Technology*, vol. 35, no. 4, pp. 161–179, 2010.

[5] S. Baxter, P. Enderby, P. Evans, and S. Judge, "Barriers and facilitators to the use of high-technology augmentative and alternative communication devices: a systematic review and qualitative synthesis," *International Journal of Language & Communication Disorders*, vol. 47, no. 2, 2012.

[6] D. J. Higginbotham, H. Shane, S. Russell, and K. Caves, "Access to AAC: Present, past, and future," *Augmentative and alternative communication*, vol. 23, no. 3, 2007.

[7] H. C. Shane, S. Blackstone, G. Vanderheiden, M. Williams, and F. DeRuyter, "Using AAC technology to access the world," *Assistive technology*, vol. 24, no. 1, 2012.

[8] P. Ola Kristensson and T. Müllners, "Design and Analysis of Intelligent Text Entry Systems with Function Structure Models and Envelope Analysis," *Analysis*, vol. 12, 2021.

[9] J. Todman, N. Alm, J. Higginbotham, and P. File, "Whole Utterance Approaches in AAC," *Augmentative and Alternative Communication*, vol. 24, no. 3, 2008.

[10] J. Shen, B. Yang, J. J. Dudley, and P. O. Kristensson, "Kwickchat: A multi-turn dialogue system for AAC using context-aware sentence generation by bag-of-keywords," in *27th International Conference on Intelligent User Interfaces*, 2022.

[11] J. Adhikary and K. Vertanen, "Language model personalization for improved touchscreen typing," in *Proc. Interspeech'23*, 2023.

[12] V. Dhakal, A. M. Feit, P. O. Kristensson, and A. Oulasvirta, "Observations on typing from 136 million keystrokes," in *Proc. 2018 CHI conference on human factors in computing systems*, 2018.

[13] P. O. Kristensson, J. Lilley, R. Black, and A. Waller, "A design engineering approach for quantitatively exploring context-aware sentence retrieval for nonspeaking individuals with motor disabilities," in *CHI Conf. on Human Factors in Comp. Sys.*, 2020.

[14] S. Cai, S. Venugopalan, K. Tomanek, A. Narayanan, M. Morris, and M. Brenner, "Context-aware abbreviation expansion using large language models," in *Proc. 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, July 2022, pp. 1261–1275.

[15] H. Curtis, T. Neate, and C. Vazquez Gonzalez, "State of the art in AAC: A systematic review and taxonomy," in *Proc. 24th Int. ACM SIGACCESS Conference on Computers and Accessibility*, 2022.

[16] L. Hao, S. Goetze, and M. Hawley, "Message recommendation strategies for tailoring health information to promote physical activities," in *HCI International 2023*, Q. Gao, J. Zhou, V. G. Duffy, M. Antona, and C. Stephanidis, Eds. Springer, 2023, pp. 536–555.

[17] L. Hao, S. Goetze, T. Alessa, and M. Hawley, "Effectiveness of computer tailored health communication on physical activity promotion for people with or at risk of long-term conditions: systematic review and meta-analysis," *J. Medical Internet Research (JMIR)*, vol. 5, 2023.

[18] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019.

[19] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A robustly optimized BERT pre-training approach with post-training," in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, Aug. 2021.

[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, 2019.

[21] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *Proc. ICASSP 2023*, 2023.

[22] S. A. Chowdhury, N. Durrani, and A. Ali, "What do end-to-end speech models learn about speaker, language and channel information? A layer-wise and neuron-level analysis," *Computer Speech & Language*, vol. 83, p. 101539, 2023.

[23] K. Vertanen and P. O. Kristensson, "The imagination of crowds: conversational AAC language modeling using crowdsourcing and large data sources," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.

[24] H. Yusufali, S. Goetze, and R. K. Moore, "Bridging the Communication Rate Gap: Enhancing Text Input for Augmentative and Alternative Communication (AAC)," in *HCI International 2023*, Q. Gao, J. Zhou, V. G. Duffy, M. Antona, and C. Stephanidis, Eds. Springer Nature Switzerland, 2023, pp. 434–452.

[25] M. Davies, "The TV and Movies corpora: Design, construction, and use," *International Journal of Corpus Linguistics*, vol. 26, no. 1, 2021.

[26] A. N. Medvedev, R. Lambiotte, and J.-C. Delvenne, "The anatomy of reddit: An overview of academic research," *Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches 10*, 2019.

[27] J. Wallat, J. Singh, and A. Anand, "BERTnesia: Investigating the capture and forgetting of knowledge in BERT," in *Proc. 3rd BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Nov 2020.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. 40th annual meeting of the Association for Comput. Linguistics*, 2002.

[29] I. Hameister and L. Nickels, "The cat in the tree–using picture descriptions to inform our understanding of conceptualisation in aphasia," *Language, Cognition and Neuroscience*, vol. 33, no. 10, 2018.

[30] J. Millet and E. Dunbar, "Do self-supervised speech models develop human-like perception biases?" in *Proc. 60th Annual Meeting of the Association for Computat. Linguistics*, 2022.