Towards Generation of Personalised Health Intervention Messages

Clara Wan Ching Ho[1,2], **Volha Petukhova**[1]
[1]Spoken Language Systems Group, Saarland University, Saarbrücken, Germany
[2] Goethe University Frankfurt, Frankfurt am Main, Germany
c.ho@ub.uni-frankfurt.de; v.petukhova@lsv.uni-saarland.de

Abstract

Self-care is essential in managing chronic diseases when patients could not always be monitored by medical staff. It therefore fills in the gap to provide patients with advice in improving their conditions in day-to-day practices. However, effectiveness of intervention messages in encouraging healthy behaviour is limited, as they are often delivered in the same manner for patients regardless of their demographics, personality and individual preferences. In this paper, we propose strategies to generate personalized health intervention messages departing from assumptions made by theories of social cognition and learning, planned behaviour and information processing. The main task is then defined as a personalised argument generation task. Specifically, an existing well-performing Natural Language Generation (NLG) pipeline model is extended to modulate linguistic features by ranking messages generated based on individuals' predicted preferences for persuasive arguments. Results show that the model is capable of generating diverse intervention messages while preserving the original intended meaning. The modulated interventions were approved by human evaluators as being more understandable and maintaining the same level of convincingness as human-written texts. However, the generated personalised interventions did not show significant improvements in the power to change health-related attitudes and/or behaviour compared to their non-personalised counterpart. Data collected for the model's training was rather limited in size and variation though.

Keywords: personalised medicine, health messages generation, content adaptation

1. Introduction

In the context of the global aging population and persistent pressure on healthcare providers to lower their service costs, self-care eHealth services that provide health interventions[1] increasingly gaining popularity. Offered health interventions often have however limited effects on patient motivation, therapy compliance and behaviour or attitude change; a personalised approach is necessary (Adler et al., 2016; Kee et al., 2018). The need for personalisation comes from two primary sources that are not necessarily exclusive: gaps in medical and personal knowledge (Rojas, 2021). *Medical knowledge* of patients may be insufficient to understand health intervention texts. Walsh and Volsko (2008) showed that internet-based consumer-health information articles were written above the recommended reading levels for the average adult. *Personal knowledge* of doctors means that they may be not aware of individual preferences, emotional state, social status and lifestyle of their patients. Knowing certain patient characteristics and preferences associated with those characteristics doctors could personalise their messages that are optimal for their patients (Kee et al., 2018).

Modern Artificial Intelligence (AI) systems enable many application scenarios which incorporate automated online interactions. Recent generative AI methods, in particular Large Language Models (LLM), offer the possibility of building text generation agents such as ChatGPT that can provide personalised content. However, the pre-trained large models are not suitable for specific applications without explicit prompting, instructions, re-training and/or adaption to a particular domain and task.

The study presented in this paper aims first at assessment of personalisation factors that may influence the interaction quality outcome, i.e. effectiveness of intervention messages for decision-making support and high treatment adherence. Our assumptions are based on the key predictions made by established models of planned behaviour, social cognition, learning and information processing. We test these assumptions in human-based study and collect initial data to design our prediction and generation models. A pipeline model is proposed which modulates medical evidence-based arguments extracted from PubMed abstracts with respect to medical and personal knowledge of the patient. Effects of linguistic modulations are evaluated in post-test experiments where human participants rate, rank and select messages as most convincing, understandable, competent and helpful. Interaction effects between participants' personal profiles and message manipulations are assessed.

The paper is structured as follows. Section 2 reviews models of individual and social aspects of decision making and information processing. We identify factors that impact the generation of convincing personalised health interventions. Section 3 introduces related NLG work in the field of person-

[1]According to the World Health Organization (WHO, 2022), self-care interventions are evidence-based tools used to promote and maintain health, prevent disease and cope with illness outside of formal health services.

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 64–72

alisation. In Section 4, our methodology, resources and architecture design are presented. Section 5 discusses pre-testing, implementation and evaluation experiments. Section 7 summarizes our findings, discusses limitations and outlines directions for future research and development.

2. Aspects of Decision Making and Behavioural Change

In order to gain patient adherence and motivate them to change their attitude and/or behaviour, it is important to identify what communication strategies are most appealing to them. Knowing patient characteristics and preferences associated with these characteristics help constructing optimal targeted intervention messages. It has been observed that patients prefer a psycho-social model of communication compared to a biomedical model, which is more commonly used by medical personnel (Kee et al., 2018). Thus, along with truthfulness and logical coherence of the arguments proposed in health intervention messages, their effectiveness relates to individual beliefs and intervention delivery aspects. *Planned Behaviour Theory* (Ajzen, 1991) and *Social Learning Theory* (Bandura and Walters, 1977) specify factors behind intentions and decisions to change attitude and behaviour comprising (1) individual attitudes towards behaviour and its outcomes: perceived importance and perceived level of readiness; (2) perceived social norms; and (3) the individual beliefs (confidence) about abilities to perform and control behaviour and its outcome. *Elaboration Likelihood Model* (Cacioppo and Petty, 1984) explains processing of persuasive messages and factors that facilitate potential attitude change associated with them. *Stereotype Content Model* (Cuddy et al., 2008) predicts the emotions associated with social groups on perceived warmth and competence of communicated messages. Theoretical predictions made by these models equip us with initial assumptions concerning the utility of intervention messages in inducing intended potential attitude and behaviour change. Figure 1 provides an overview of the basic assumptions tested in this study.

We assume that the quality of reached outcomes in terms of therapy compliance, motivation and attitude/behaviour change will depend on the content quality of interventions and patient personal characteristics. These two major factors, in their turn, depend on the perceived levels of agency/competence and warmth/communion - the big two of social cognition (Fiske, 2018).

The level of competence of the arguments presented in intervention messages are defined in terms of: (1) quality of the information provided, e.g. level, expert language use and expressed certainty level; and (2) framing effects, e.g. presenting options in positive terms (survival rates for a procedure) or in negative terms (mortality rates for a procedure). We assume that interventions based on valid medical evidence formulated in professional, concrete and confident language, and appropriate framing effects will be perceived as competent, see also (Guenoun and Zlatev, 2023).

Personal characteristics influencing the perceived levels of competence and warmth concerns general characteristics of the communicators (i.e. power/status, gender and age) and their personality trait profile (i.e. BIG 5; (McCrae, 1992)). Certain personality traits could be associated with higher levels of perceived competence and warmth in humans and agents. In the line with Dubois et al. (2016), we expect a fit effect between levels of competence and warmth of the generated interventions and patient's preferences on outcome quality: if the competence and warmth levels match, the quality would be higher than when there is a mismatch observed. This is compliant with *Elaboration Likelihood Model* (Cacioppo and Petty, 1984), which suggests that potential attitude change in persuasion could be seen as an act of information processing determined by the use of 1) *central route* which involves more cognitive processes and elaborated processing or 2) *peripheral route* which involves heuristics and cues pickups in processing information, based on an individual's motivation and abilities. The theory states that when a person is motivated and able to process a persuasive message that reinforces one's attitude, with a change in cognitive structure, then likely the central route would be taken, resulting in an attitude change. Otherwise, either peripheral route would be taken to process the message leading to a potential attitude shift temporarily, or there would not be an attitude change.

3. Related Work on Generation of Natural Language Interventions

Reiter and Dale (1997) proposed a classical NLG pipeline model that has been widely used and modified to suit a range of purposes, generating texts from an abstract goal to specific wordings. The model includes three components: a *Text Planner*, a *Sentence Planner* and a *Linguistic Realiser*.

A more recent modification proposed by Pauws et al. (2019) adapted the data-to-text architecture for medical domain application. Medical reports are generated automatically from patient's clinical outcomes. Other than the three components in the classical pipeline, another layer of data analysis before content determination is added, allowing output to contain different messages according to one's clinical outcomes. Here, medical knowledge

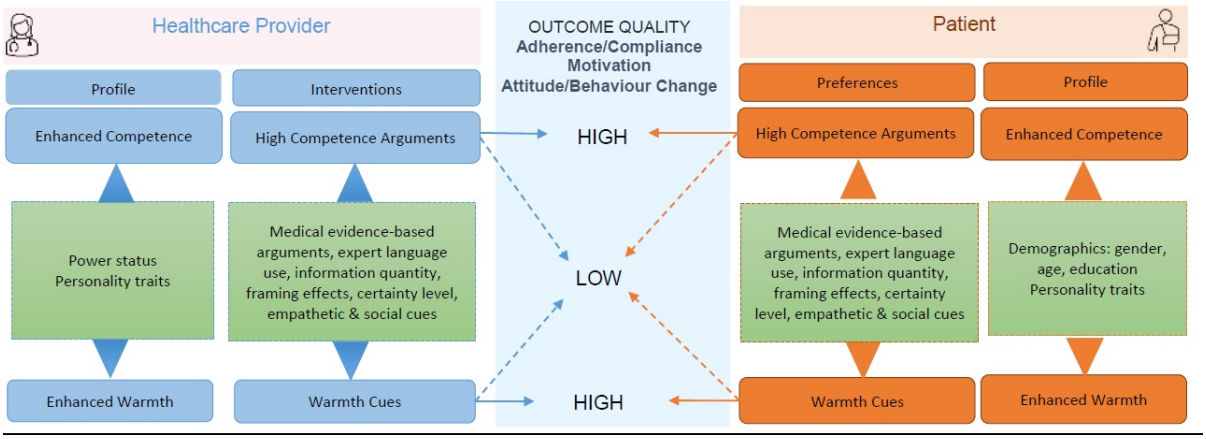

Figure 1: Overview of the key variables and their predicted interplay for health interventions.

was integrated. For example, when one's blood sugar is higher than a threshold as in the knowledge base, a warning message is generated and included in the report.

Mairesse and Walker (2007) implemented a model based on the classical pipeline model of Reiter and Dale (1997) to generate dialogues that mimic people with different levels of extroversion. Language cues for extroverted and introverted people, for example the frequently occurred negations, were considered. Parameters included, but were not limited to, self-references, lexicon frequency, hedge variation and concessions polarity. Two generation approaches were applied: (1) dialogues were generated based on the hypothesised parameters from previous studies, and (2) over-generating dialogues and selecting one that is the most similar to the target level of extroversion. Outputs were evaluated by human raters in terms of their perceived level of extroversion.

Guenoun and Zlatev (2023) compiled a list of linguistic cues as signals of competence and warmth. Their analysis showed that the use of infinitive verbs, nouns and determiners can be considered as accurate signals to quantify competence appeal, and the use of personal pronouns, verbs and wh-determiners as variables to quantify warmth appeal.

We follow the classical pipeline model of Reiter and Dale (1997), taking the approach of Pauws et al. (2019) integrating medical domain knowledge for the persuasive 'competent' content, and the over-generating and matching style applying regression approaches as by Mairesse and Walker (2007). Features studied by Guenoun and Zlatev (2023) are incorporated to quantify perceived competence and warmth appeals.

4. Methodology

The domain selected for our use case concerns the treatment of diabetes. To generate health interventions, *data* were collected from PubMed abstracts and reports of the American Diabetes Association with reference to PubMed publications[2]. Data was manually segmented into an argument structure (Mayer et al., 2020), and used for further personalisation. For this, a *pre-test* was designed based on known persuasive strategies, personality traits and linguistic features.

In a pre-test, demographic and personality profiles of respondents were collected, along with their judgements of manually modified texts to assess our initial intuition on persuasiveness, understandability, perceived professionality and perceived helpfulness. Correlations between respondent's personal profiles and linguistic preferences inferred from their judgements were analysed. Discovered effects were considered as parameters predicting one's preferred linguistic delivery of a persuasive intervention.

A pipeline generation model has been proposed to rely on the predicted linguistic parameters related to the perceived competence and warmth. Extracted evidence-based arguments were enriched with alternative medical terms and their definitions from the Unified Medical Language System (UMLS) term bank (Bodenreider, 2004).

Finally, the quality of modulated arguments incorporated into personalised interventions were automatically evaluated and assessed in a post-test by human evaluators.

4.1. Data and Pre-processing

Medical claims related to self-management actions were extracted, see Table 1 for an example. We assumed that (pre-)diabetic conditions, treatments and prevention measures are publicly relatively well known. According to the Centers for Disease Control and Prevention (2022) large portion of Western population suffers from diabetes, knows somebody

[2] diabetesjournals.org

Type	Content
Major claim	You should minimise alcohol intake.
Claim (support)	Minimal alcohol intake lowers health risk for people with diabetes.
Premise (support)	Alcohol intake may place people with diabetes at increased risk for delayed hypoglycemia.

Table 1: Example of a PubMed excerpt as an argument structure of Mayer et al. (2020).

in their family or close social group diagnosed with it or thinks to have sufficient knowledge about the disease. For instance, many studies report that the majority of respondents (up to 97.3% in Italy) had heard about diabetes (Pelullo et al., 2019). Thus, initial attitudes and respective potential changes can be tested rather reliably when assessing the effectiveness of the generated interventions.

From PubMed abstracts of Randomised Controlled Trials (RCT), 16 major claims concerning treatment or life quality improvement actions were selected. Excerpts were manually segmented at clause boundaries, to fit into the argument structure. Table 1 illustrates claims and evidence (premise) to persuade people to *"Minimise alcohol intake"*. For each major claim supporting and attacking claims and at least one premise were generated resulting in 32 claims and 35 premises in total.

5. Experimental Design

5.1. Pretest

A pre-test is conducted to collect preferences of respondents with various demographics for different linguistic delivery styles and to test initial assumptions that modulations of linguistic features: (1) are acknowledged by respondents; (2) have effects as predicted by theoretical models; and (3) lead to attitude change.

Data: Ten claims were selected for pre-testing: five are supporting the major claim, and other five are attacking those. Six variables, known from previous research, were selected for linguistic modulations and concern *Appeal* (competence/warmth), *Text length* (long/short), *Framing* (risk/benefit), *Lexical complexity* (complex/simple), *Concreteness* (numbers/textual delivery) and *Grammatical voice* (passive/active).

The tested claims were edited manually removing redundancy and generating the missing either attacking or supporting counterpart. This resulted in 12 intervention pairs, where in each pair only one linguistic variable is modulated.

Questionnaire has been designed comprising five parts to collect *participants profile*, to assess *pre-intervention attitudes*, to rate *individual interventions*, to compare *pairs of interventions*, and to identify *post-intervention attitudinal change* if any.

To design participant's profile, information about one's knowledge/experience with diabetes, gender, age and highest attained education level were collected. Further, participants were asked to complete an online Big Five Personality Test of Open-Source Psychometrics Project[3] Personality profiles corresponding to *extroversion*, *neuroticism*, *agreeableness*, *conscientiousness* and *openness* were computed.

To assess pre- and post-intervention attitudes, respondents were asked to rate on a 7-point Likert scale ten actions that have potentials in improving one's diabetic conditions.

To assess intervention arguments, respondents were presented one major claim together with a relevant base claim (either attacking or supporting) and a premise, and asked to rate them on how much they agree that the arguments are understandable, helpful, professional and persuasive (7-point Likert scale). In pairwise comparison, respondents were given one major claim and a pair of modulated premises and asked to select one which fits the best the perceived level of the tested effects, e.g. perceived helpfulness.

Results: 32 respondents participated in experiments, all English non-native speakers; 58.1% of respondents were female and 38.7% were male; all respondents have at least heard of diabetes as a medical condition; about half of the respondents were between 16 and 30 years old, 32.3% of them between 46 and 60 years old, and 12.9% between 31-45 years old; 90% of the respondents have received at least one bachelor's degree and over 30% had received at least one postgraduate degree.

The pre-test data has provided useful insights showing that the tested linguistic modulations were perceived by respondents as intended, and can be modelled as parameters in personalised intervention generation. However, it was concluded that a pairwise simple correlation between individual linguistic variable and profile factor is not sufficient to adequately quantify targeted modulation extents. Instead, the interplay between factors should be taken into account when implementing the personalisation model and therefore have contributed to the choice of incorporating random forest models in the pipeline model.

5.2. Pipeline Model

The pipeline has two streams, one dealing with the linguistic content (referred to as NLG Stream), the other dealing with the user's personal profile

[3] https://openpsychometrics.org/tests/IPIP-BFFM/

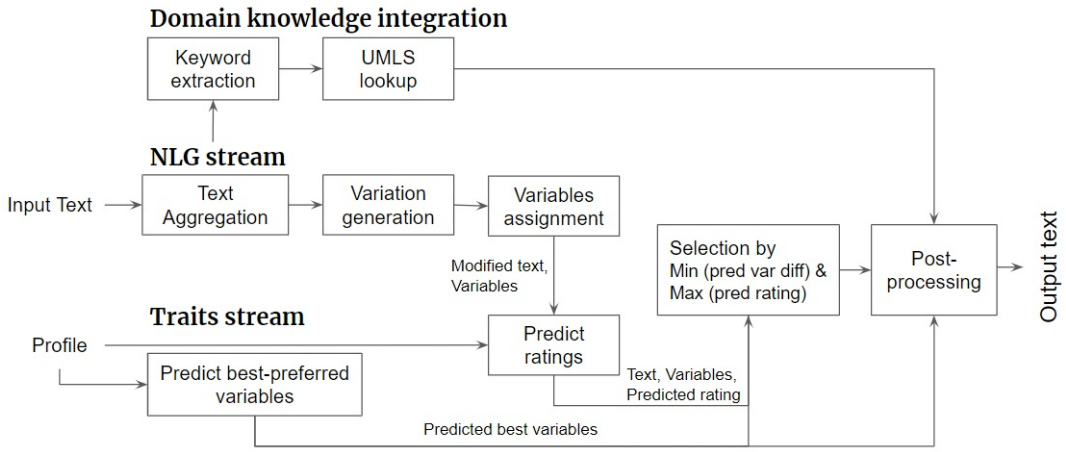

Figure 2: Architecture of the proposed pipeline model.

Parameter	Variables
Appeal to Competence	Average of: VBN tokens / total tokens NN tokens / total tokens DT tokens / total tokens
Appeal to Warmth	Average of: PRP tokens / total tokens VB tokens / total tokens WDT tokens / total tokens
Numeric delivery	local numeric token / max numeric token
Text length	local token count / max token count
Lexical complexity	local average token length / max average toke

Table 2: Parameters to modulate linguistic features. VBN stands for Verb, past participle; NN for Noun, singular or mass; DT for Determiner; PRP for Personal pronoun; VB for Verb, base form; WDT for Wh-determiner.

(referred to as Traits Stream). Additionally, medical domain knowledge (i.e. UMLS) is consulted to look up definitions of medical terms. Decisions to integrate the definition to augment an intervention argument is made at the post-processing step.

In the Traits Stream, information of the user's profile including age, gender and personality traits scores (Goldberg, 1993) serves as input. Preferences for designated linguistic variables are predicted by two Random Forest regression models. Those are trained on the pre-test data where one model predicts a rating given an individual's profile and linguistic features, and the other model predicts and ranks linguistic features given an individual's profile. As a result, weights are assigned to respective linguistic features in an intervention argument and passed for comparison with weights of the generated options in the NLG Stream.

In the NLG Stream, an excerpt of the same major claim and premise serve as input. They are processed by the sentence aggregation component based on the BART paraphrase model (Lewis et al., 2019) which generates interventions of different lengths with minimal lexical or syntactic changes, and redundant content removed. Repetitive interventions generated are filtered out by Levenshtein distance. The selected interventions are paraphrased with T5 paraphrase model PARROT (Damodaran, 2021). In this way, interventions with a great diversity in lexical, syntactic and potentially semantic changes are (over-)generated. The linguistic variables of the over-generated intervention arguments values are assigned, compared with the predicted values of the corresponding arguments of the Trait Stream and ranked. The best matching intervention, i.e. one with the highest predicted rating and minimal difference between the variables in linguistic features of the predicted preference and generated options, is selected for generation and returned to the user.

5.2.1. Linguistic Features for Personalisation

To quantify linguistic features of the generated interventions, five parameters were considered: *appeal to competence*, *appeal to warmth*, *relative level of numeric delivery*, *relative text length* and *relative lexical complexity*, see Table 2. With reference to the output of paraphrase generation, values for the variables were assigned in batches. A batch is the set of paraphrases generated from the same intervention claim or premise. For each batch, the maximal counts of numeric tokens, maximal token count and maximal average token length were computed. The model looped through all entries in the same batch and divided the local counts by the computed maximal counts to assign their relative level of numeric delivery, relative text length and relative lexical complexity, resulting in five values of the five parameters, each between 0 and 1.

To estimate appeal to competence and appeal to warmth, an average of three local variables

as listed in Table 2 were considered. For this, Part-of-Speech (POS) tagging with Python Natural Language Toolkit (nltk) library was performed, and count estimates were computed as explained above. Values of the five parameters, along with a participant's profile, were sent to a Random Forest model to predict a convincingness score for each of the intervention generated.

5.2.2. Domain Knowledge Integration

The domain knowledge integration component essentially identifies the key medical concepts in the text and looks up definitions for the respective term in knowledge base. Keywords and phrases were extracted as candidates using KeyBERT model (Grootendorst, 2020) from a pypi package. Subsequently, terms were queried in the Consumer Health Vocabulary (CHV) term bank with UMLS API. Given that the CHV is a medical term bank of common medical terms, if a term was found in CHV, it was considered unnecessary to provide readers with additional information about the term as it is already commonly known. The remaining terms were queried with the UMLS API in the available English medical term banks and the term entry with its respective definitions were retrieved. For simplicity, only the first matching entry was returned. The list of filtered terms, their first matching term entry and their respective definitions were passed on to the final post-processing component for rule-based term-replacement after suitable intervention arguments were generated.

6. Evaluation

Intervention Quality as Texts The quality of generated interventions was assessed automatically based on cosine similarity and well-formedness. While cosine similarity assesses the degree of semantic information retained in the modulated intervention message, well-formedness assesses its grammaticality.

An off-the-shelf similarity model from sentence transformers (Reimers and Gurevych, 2019) was used. Semantic similarity scores ranging from 0.8 to 0.97 were obtained for all generated intervention messages when compared to their unmodified counterparts. These values indicated that the modulated messages largely retained the information of the original interventions.

The well-formedness was automatically assessed with the off-the-shelf model of Kumar (2020). The unmodified interventions got a mean well-formedness score of 0.5, with a range of approximately from 0.3 to 0.65. The generated modulated interventions exhibited larger variations, where their well-formedness ranged between 0.1 and 0.9. Two-thirds of modulated interventions had a lower average score than the unmodified ones. 60% of the modulated intervention messages were within the range of ± 10% of the mean of the unmodified ones. Figure 3 summarises the results.

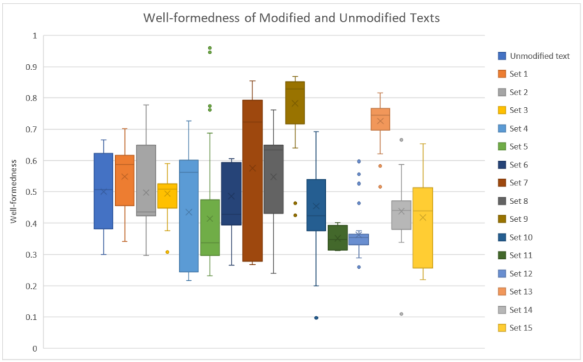

Figure 3: Mean well-formedness of unmodified and 15 sets of modified intervention messages.

Association with Decision Making Aspects The pipeline model did not specify personalisation strategies in linguistic features modulation, instead interventions were generated by selecting preferences expressed in linguistic cues relevant for decision making aspect. Based on the in-depth analysis and achieved effects such strategies can be defined. The following samples of interventions demonstrate how different syntactic structures, subjectivity, mood, information load and vocabulary use can be associated with perceptions triggered by generated interventions.

From pre-defined prompts of "You should [action]" for major claims, four patterns of subjectivity expressions were observed in modified intervention messages:

(i) identical to the original input of "You should [action]";

(ii) "I recommend you [action]". Both (i) and (ii) display a higher level of subjectivity and start with a personal pronoun, where (i) is stronger in tone than (ii);

(iii) "It is recommended to [action]" is seen as objective and neutral; and

(iv) "The best ... is [recommendation]", paraphrases statements with a recommended action in replacement of a potentially harmful one.

Modified prompts which differ in subjectivity are presented in (1):

(1) You should lose some weight.
I recommend you lose weight.
It is recommended to extend the time spent sleeping.
The best replacement for sugar-sweetened drinks is water.

Apart from subjectivity, generated interventions differed in mood, including indicative, imperative, conditional and interrogative as exemplified in (2) .

(2) It is important to do resistance training and aerobic exercises. (indicative)
Take zinc supplements to slow the development of diabetes. (imperative)
If you want your health to improve you should take supplements that contain b12. (conditional)
Do you have to do a balance exercise? Short-term structured strength and balance training did not affect HRQoL; there were no significant differences between groups on the primary outcomes of PCS score and EQ-5D-5L index score. (interrogative)

Modified intervention messages A and B below in (3) show how the diversity in language cues may encourage intended attitude hence behavioural change.

(3) **Unmodified Intervention Message**
You should minimize alcohol intake. Minimal alcohol intake lowers health risk for people with diabetes. Alcohol intake may place people with diabetes at increased risk for delayed hypoglycemia. This is particularly relevant for those using insulin or insulin secretagogues who can experience delayed nocturnal or fasting hypoglycemia after evening alcohol consumption.

Modified Intervention Message A
Reduce the quantity of alcohol. Recommended for those using insulin or insulin secretagogues who experience delayed nocturnal or fasting Hypoglycemia (Abnormally low level of glucose in the blood) after evening alcohol consumption.

Modified Intervention Message B
It is important that you cut down on your alcohol consumption. This is particularly relevant for those using insulin or insulin secretagogues that may experience delayed nocturnal or fasting Hypoglycemia (Abnormally low level of glucose in the blood) after evening alcohol consumption.

Imperative and indicative moods are observed respectively at sentence beginnings of the two messages. The imperative mood in A conveys a sense of certainty and urgency, relevant to a higher level of perceived readiness according to the Planned Behaviour Theory (Ajzen, 1991). The expression "it is important that you" in B is related to the increase of perceived importance in the aspect of behavioural intention. Intervention Message B signals closeness with addressees when using personal pronouns, encouraging an in-group association of the warmth appeal in the Stereotype Content Model (Cuddy et al., 2008).

Diversity in vocabulary use is observed, such as the replacement of "minimize" to "cut down on" and "reduce". They can be seen as presentations that are more or less colloquial, establishing different levels of closeness, which is relevant for the competence/warmth appeal.

If parameters are set correctly, the model can personalise texts with a great variation in linguistic features to closely match individual linguistic preferences or targeted perception effects.

6.1. Post-test

To assess the intended effects of personalisation, understandability, likeability and convincingness and the quality of the generated interventions, a post-test has been conducted. We also assessed naturalness, perceived redundancy and coherence of the generated messages.

Data From the 16 major claims presented earlier, 15 were selected for the post-test: eight expressing the supporting stance, and the other seven the attacking stance. The base claims were used as unmodified interventions and proposed for personalisation.

Three types of interventions were tested: (1) unmodified arguments from medical excerpts; (2) the best matching intervention generated by the pipeline model and matching the individual preferences; and (3) the worst matching intervention generated by the pipeline model and selecting the least matching individual profile. Note that the best matching and worst matching interventions vary for each participant as they were generated based on their individual profiles.

Five parameters (appeal to competence, appeal to warmth, relative level of numeric delivery, relative text length and relative lexical complexity) were modulated.

Questionnaire Design The post-test was distributed as a questionnaire with two parts: (1) collection of participants' profiles in terms of their demographics and personality traits and is identical to that of the pre-test; and (2) a total of 15 sets of personalised and at least one non-personalised interventions were *ranked* and *rated* on a 7-point Likert scale. Additionally, the level of *information retainment* was assessed. Three randomly selected unmodified, best and worst matching intervention messages were evaluated on criteria of well-formedness (coherent and natural), understandability, redundancy and likeability (convincing). The later criteria were meant to test some of the study's hypotheses in the perception of personalised linguistic delivery.

Results 21 respondents participated in the study. All respondents were required to have at least heard of diabetes as a medical condition and have not

Evaluation Criterion	Preference Matching Setting		
	Best	**Worst**	**Unmodified**
Text Quality Evaluation			
Coherence	4.73* (± 1.5)	4.70* (± 1.6)	5.67 (± 1.5)
Naturalness	4.36* (± 0.2)	3.93* (± 0.1)	5.83 (± 0.3)
Redundancy	2.20* (± 1.6)	2.53* (± 1.5)	3.17 (± 1.7)
Perception Evaluation			
Likeability	4.37 (± 1.7)	4.31 (± 1.4)	4.53 (± 1.6)
Understandability	4.86* (± 1.9)	5.27 (± 1.7)	4.33 (± 1.7)
Convincingness	4.73 (± 1.5)	4.70 (± 1.5)	5.67 (± 1.8)

Table 3: Overview of the average ratings in text quality and perception evaluation experiments on the 7-point Likert scale. * marks statistically significant differences when compared to an unmodified intervention argument. Standard deviation is provided in brackets.

participated in the pre-test. 15 respondents successfully completed both parts of the questionnaire, all of them were between 16 and 30 years old, with 53.3% male and 46.7% female; 93.3% of the respondents have received at least one bachelor's degree and 13.3% have received a master's degree.

The results showed that the generated interventions were rated as more understandable than unmodified ones, see Table 3. This is most probably due to the simplification and added definitions of medical terms. Results show statistically significant differences where the best matching rated approximately 30% higher than unmodified ones in understandability (p=0.049).

Likeability fluctuates between test sets (Figure 4), which may be a result of the instability in paraphrasing quality. Nevertheless, results show that the likeability of the generated interventions is at least competitive with that of the unmodified arguments. The mean ratings of the five sets show that the three types of interventions were rated similarly in terms of likeability, where the unmodified ones receive the highest and the worst matching ones the lowest scores.

In both rating in ranking tasks, unmodified interventions are rated as the most redundant and best matching texts are the least redundant.

The rated naturalness and coherence of the modified interventions are noticeably lower than human-written unmodified texts. The results are understandable given the lack of grammatical and semantic check in selection of paraphrases. There are no statistically significant differences observed in convincingness of generated modulated and unmodified interventions, suggesting that the automatically generated messages are at least not less convincing after the performed modulations.

According to the post-test results, interventions generated by the model are in general less redundant, more understandable and as likeable and convincing as the unmodified arguments. However they are less natural and potentially less coherent.

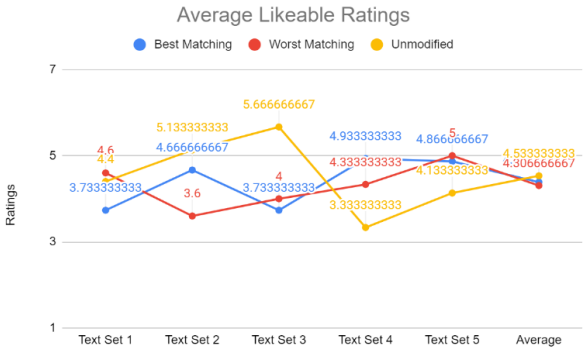

Figure 4: Average likeability ratings.

7. Discussion and Conclusion

This study evaluated the argument generation approach for medical domain application in personalising intervention messages. A pipeline model was implemented to process health interventions containing medical evidence based arguments and convert them into personalised health intervention messages. Medical domain knowledge is integrated to simplify and explain medical terms for higher understandability.

The implemented model was evaluated and produced good quality health interventions. Despite perceived as less natural, modulated interventions were rated by human evaluators as likeable and convincing as the unmodulated ones, while performing better on understandability and conciseness criteria.

Modulated interventions exhibited a high diversity in lexical and syntactic structures given different profiles, which potentially can be used to specify various persuasion strategies. Currently, no module in the model that defines or selects persuasive strategies is designed.

Further work is required to improve system's personalisation capabilities. Personalisation factors are numerous and show complex interplay, these additional effects need to be analysed in a more detailed study with higher number of participants of various demographics, personalities and emotional states. Real patient data recorded in authentic doctor-patient communicative settings is ideal but hard to obtain. We, therefore, opt for better patient simulations and simulations of interactive situations which will allow better experimental control.

8. Acknowledgments

The authors are also very thankful to anonymous reviewers for their valuable comments.

9. Bibliographical References

Rachel F Adler, Francisco Iacobelli, and Yehuda Gutstein. 2016. Are you convinced? a wizard of oz study to test emotional vs. rational persuasion strategies in dialogues. *Computers in Human Behavior*, 57:75–81.

Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.

Albert Bandura and Richard H Walters. 1977. *Social learning theory*, volume 1. Englewood cliffs Prentice Hall.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

John T Cacioppo and Richard E Petty. 1984. The elaboration likelihood model of persuasion. *ACR North American Advances*.

Centers for Disease Control and Prevention. 2022. National diabetes statistics report. Technical report.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for nlu.

David Dubois, Derek D Rucker, and Adam D Galinsky. 2016. Dynamics of communicator and audience power: The persuasiveness of competence versus warmth. *Journal of Consumer Research*, 43(1):68–85.

Susan T Fiske. 2018. Stereotype content: Warmth and competence endure. *Current directions in psychological science*, 27(2):67–73.

Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*, 48(1):26.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Bushra S Guenoun and Julian J Zlatev. 2023. Sending signals: Strategic displays of warmth and competence. *Working Paper 23-051*.

Janine WY Kee, Hwee Sing Khoo, Issac Lim, and Mervyn YH Koh. 2018. Communication skill in patient-doctor interactions: learning from patient complaints. *Health professions education*, 4(2):97–106.

Ashish Kumar. 2020. Query wellformedness scoring.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

François Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 496–503.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press.

RR McCrae. 1992. Revised neo personality inventory (neo-pi-r) and neo five-factor inventory (neo-ffi) manual. *Psychological Assessment Resources. Odessa, FL*.

Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter. 2019. Making effective use of healthcare data using data-to-text technology. In *Data Science for Healthcare*, pages 119–145. Springer.

Concetta P Pelullo, Riccardo Rossiello, Roberto Nappi, Francesco Napolitano, Gabriella Di Giuseppe, et al. 2019. Diabetes prevention: knowledge and perception of risk among italian population. *BioMed research international*, 2019.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Julian Andres Ramos Rojas. 2021. *Exploring AI-based personalization of a mobile health intervention and its effects on behavior change, motivation, and adherence*. Ph.D. thesis, Carnegie Mellon University Pittsburgh, PA.

Tiffany M Walsh and Teresa A Volsko. 2008. Readability assessment of internet-based consumer health information. *Respiratory care*, 53(10):1310–1315.

WHO. 2022. Self-care interventions for health.