Universität Zürich[UZH]

Master's thesis
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of
Master of Arts

Respond in Style: Personalising Review Response Generation by Optimising Continuous Prefixes

Author: Helen Schaller
Student ID Nr.: 16-731-010

Examiner: Prof. Dr. Martin Volk
Supervisor: Tannon Kew

Department of Computational Linguistics

Submission date: 29.11.2022

Abstract

Responses to online reviews need to be appropriate not only in terms of content but also style. Previous work on the automatic generation of review responses using sequence-to-sequence modelling has neglected the task of personalising the responses to individual businesses despite the importance of styling responses for corporate identity construction.

A naive approach to achieve personalisation in review response generation would be to fully fine-tune as many large pre-trained language models as there are companies. However, this approach turns prohibitively memory-intensive with a growing number of businesses.

In this master thesis, we compare a lightweight alternative called prefix-tuning [Li and Liang, 2021] to full fine-tuning for generating general as well as personalised review responses. Prefix-tuning optimises task-specific continuous prefixes that are prepended to the model while the pre-trained language model's weights are left unchanged.

We show that prefix-tuning approximates the performance of fine-tuning when adapting a pre-trained model to the downstream task of review response generation. In terms of generating styled review responses, the prefix-tuned model produces more personalised as well as diverse responses than the fine-tuned model when combined with a sampling-based decoding approach.

Zusammenfassung

Antworten auf Online-Kundenrezensionen müssen nicht nur inhaltlichen, sondern auch stilistischen Anforderungen entsprechen. In früheren Forschungsarbeiten über die automatische Antwortgenerierung zu Kundenrezensionen mithilfe von Sequence-to-Sequence-Modellierung wurde der stilistische Aspekt nicht berücksichtigt, obwohl der Antwortstil eines Betriebs einen substanziellen Bestandteil des Unternehmensauftritts darstellt.

Ein naheliegender Ansatz zur Generierung von personalisierten Antworten auf Kundenrezensionen wäre das Finetuning von ebenso vielen vortrainierten grossen Sprachmodellen (*language models*) wie die Anzahl von Unternehmen. Mit einer wachsenden Zahl von Betrieben würde dieser Ansatz jedoch schnell speicherintensiv werden.

In dieser Masterarbeit vergleichen wir Prefixtuning [Li and Liang, 2021], eine verschlankte Form von Finetuning, mit dem traditionellen Finetuning. Wir wenden beide Trainingsmethoden auf zwei Problemstellungen an: der automatischen Generierung von generellen wie auch personalisierten Antworten auf Kundenrezensionen. Während des Prefixtunings werden die Parameter des vortrainierten Sprachmodells nicht verändert. Stattdessen werden kontinuierliche Präfixe optimiert, die an das Sprachmodell angehängt werden, wobei ein Präfix pro Aufgabe erstellt wird.

Wir belegen, dass Prefixtuning die Leistung von Finetuning approximiert, wenn es darum geht, ein generelles Sprachmodell an die Aufgabe anzupassen, Kundenrezensionen zu beantworten. Besteht die Aufgabe darin, die Antworten an die Stile unterschiedlicher Unternehmen anzupassen, übertrifft das mit Prefixtuning optimierte Sprachmodell jenes, welches durch Finetuning trainiert wurde: In Kombination mit einem Sampling-basierten Dekodierungsverfahren produziert es sowohl stärker personalisierte als auch diversere Antworten.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor Tannon Kew for his constructive advice, valuable feedback, and consistent support throughout my thesis.

Thanks also goes to the researchers from the ReAdvisor project for so kindly sharing their data, the greeting and salutation removal model, and their evaluation scripts with me.

I would also like to extend my sincere gratitude to Prof. Dr. Martin Volk who sparked my passion for programming and computational linguistics from the very first lecture, and the entire staff of the Department of Computational Linguistics at the University of Zurich who shared their knowledge with so much enthusiasm.

I thank my parents from the bottom of my heart for supporting me throughout my studies. I also thank Kevin with all my heart for his unwavering support and faith in me, and for proofreading this thesis and sharing his helpful feedback.

Thanks to all of you for nurturing my knowledge, assisting my work, and encouraging my progress. Without your contributions, this would not have been possible.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
BPE	Byte-Pair Encoding
BS	Beam Search
eWOM	Electronic Word of Mouth
FFNN	Feedforward Neural Network
FT	Fine-Tuned
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
IR	Information Retrieval
ITF	Inverse Term Frequency
LM	Language Model
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
PER-PT	Personalised Prefix-Tuned
PT	Prefix-Tuned
RNN	Recurrent Neural Network
RR	Review Response
RRGen	Review Response Generation
RQ	Research Question
Seq2seq	Sequence-to-Sequence
TF–IDF	Term Frequency—Inverse Document Frequency

1 Introduction

1.1 Motivation

With the rise of the internet, more and more people share their thoughts and opinions not only with the circle of acquaintances in their vicinity but also with people all over the world. The new global reach of online experience sharing has not only a social impact on all those participating in this online exchange of views but also an economical consequence, as consumer experiences start to be narrated in platforms on the web. This especially affects the tourism industry, as information about services and goods in the holiday destination are not readily accessible via the people living in one's proximity. Information obtained online replaces traditional word-of-mouth advertising in this context.

Internet platforms such as TripAdvisor[1] are popular for sharing and obtaining information of experiences with hospitality businesses. This openly accessible information forms what is called the "electronic word-of-mouth" ("eWOM") [Katsiuba et al., 2022; Zhang and Vásquez, 2014]. What consumers claim about a company online influences fellow customers' decisions and thus spendings, as they consider others' opinions in their decision-making process [Zhang and Vásquez, 2014]. However, companies are not purely at the mercy of online reviewers' judgements: A study conducted by TripAdvisor's research team has shown that responding to an online review appropriately improves a customer's probability of recommending the business by at least 20% [Barsky and Frame, 2009]. Moreover, responding to customer reviews on TripAdvisor has a positive effect on future ratings, both in terms of improved ratings and increased review volume [Proserpio and Zervas, 2017; Xie et al., 2016]. These findings show that businesses interested in maintaining a good reputation online should invest in an effective online review management strategy. Addressing an ever-growing number of customer reviews is time-consuming, and many businesses, predominantly small ones, lack the resources to provide online feedback. (Semi-)automated review response generation has the potential to sup-

[1] `https://www.tripadvisor.com/` (last accessed 21 October 2022).

port the process of producing qualitative responses [Katsiuba et al., 2022].

Reacting to customer feedback online has not only the direct purpose of handling customer satisfaction but also of constructing a corporate identity and reputation [Creelman, 2015]. Therefore, it is required that the responses suit the style of the business in question. For example, figure 1 shows that different restaurants respond differently to the same review. To date, personalised review response generation is a research gap. Adapting automatic review response generation to individual businesses in a simple and efficient way would boost the applicability of review response generation systems immensely.

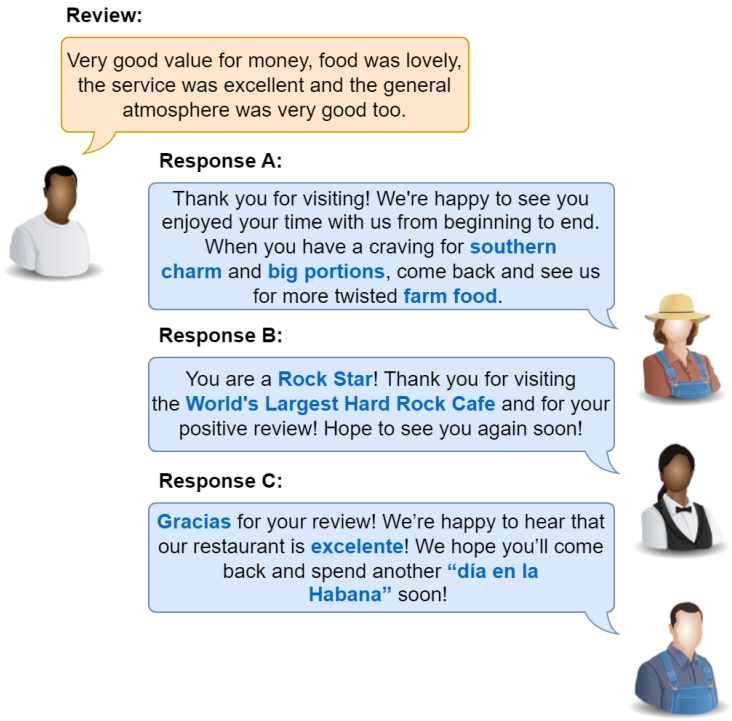


Figure 1: Personalising responses for restaurant reviews: Depending on cuisine and branding, different restaurants react differently to the same review. Differences occur both on the level of linguistic style (e.g. in the choice of Spanish words in response C) and on the level of content (focus on on hearty Southern US cuisine in response A or rock theme in response B). The review and the responses were published as they are on TripAdvisor.

1.2 Task and Research Interests

The goal of review response generation is to automatically produce review responses that are appropriate in terms of the reviews' content and helpful for improving the eWOM of a business. This is a subtask of the broader field of text generation.

Text generation typically leverages pre-trained language models that are trained on large amounts of unsupervised data in order to produce text. For task-specific text generation, these language models are conditioned to produce texts that adhere to certain criteria, such as matching certain styles, domains or tasks. For review response generation, a language model is trained to generate responses conditioned on some customer reviews.

In order to generate responses that additionally suit the style of a business, a language model is conditioned to solve the task of review response generation in the response style of a particular business. With this style specification, producing review responses for each business is an independent task. In this work, we are particularly interested in generating review responses in the style of individual restaurants.

We consider a novel approach and apply prefix-tuning [Li and Liang, 2021] to solve the task of review response generation. Li and Liang [2021] propose prefix-tuning as a lightweight alternative to the standard approach of leveraging a pre-trained model, i.e. fine-tuning. Prefix-tuning is more memory-efficient than traditional fine-tuning, as it keeps the original weights of the pre-trained model frozen while optimising only a small continuous prompt for each task. For standard fine-tuning, on the other hand, we adapt the pre-trained model's own weights which requires storing a full copy of the model for each task.

The central research questions of this thesis are the following:

RQ1. Is an efficient fine-tuning approach such as prefix-tuning a viable alternative to traditionally fine-tuning a pre-trained model for the general downstream task of review response generation?

RQ2. Given a prefix for each restaurant in the data set, how well can these prefixes capture stylistic characteristics of their respective restaurant to generate personalised review responses?

For RQ1, we first investigate if prefix-tuning is a suitable approach for adapting a pre-trained model to review response generation in general. After that, we turn to RQ2, the main focus of this thesis, where we gauge the effectiveness of prefix-tuning for personalising review responses.

Review response generation is not a trivial task. Internal knowledge about the business, world knowledge, and tact are required to create qualitative responses. While investigating the usability of prefix-tuning for adapting a pre-trained model towards producing (personalised) review responses, we also discuss the following secondary research question:

RQ3. What are the challenges in (personalised) review response generation?

For RQ1 and RQ2, we primarily focus on the results from our conducted experiments. RQ3 gives us the opportunity to examine human-written as well as generated data in order to obtain a more general overview of what makes review response generation a challenge.

1.3 Thesis Structure

In chapter 2, we first introduce previous research on review response generation, personalisation and its evaluation in the context of dialogue systems, and two methods of adapting pre-trained language models: fine-tuning and prefix-tuning. In chapter 3, we compare data sets of former studies on personalised text generation. We also list the considerations we made while compiling our own data set of TripAdvisor restaurant review-response pairs and present the main characteristics of our data set. In chapter 4, we illustrate how we adapt Li and Liang's implementation of prefix-tuning [2021] in order to generate review responses. In chapter 5, we evaluate the generated responses by the fine-tuned baseline and experimental prefix-tuned models in order to answer RQs 1 and 2. In chapter 6, we discuss the major challenges of review response generation and personalisation evaluation to shed light on RQ3, and chapter 7 summarises the findings of this work and concludes this thesis.

2 Background

In this chapter we introduce previous research related to our task, methods and evaluation metrics. More specifically, we present a small body of literature on review response generation in section 2.1. The following sections introduce methods (2.2) and evaluation metrics (2.3) for personalisation in dialogue systems. Section 2.4 compares two ways of adapting large pre-trained language models towards downstream tasks, i.e. traditional fine-tuning and prefix-tuning [Li and Liang, 2021].

2.1 Review Response Generation

The first two studies of Gao et al. [2019a] and Zhao et al. [2019] treated the topic of review response generation, although not applied to reviews in the hospitality domain. Gao et al. [2019a] extended an RNN seq2seq model such that it generated responses to app reviews from the Google Play store. Their model architecture allowed for inputting not only the review but other text-external information that is specific to app reviews, i.e. app category, keywords, and user rating. The aim of these additional cues was to help generate responses about the topics in the user reviews with an adequate sentiment. They showed in an ablation study that the additionally provided information indeed helped to improve the generated responses compared to the baseline attentional seq2seq model. Zhao et al. [2019] adapted a seq2seq model towards replying to online store reviews for clothes. Their motivation was to improve sales by writing high-quality review responses. As they noted, only a small fraction of online reviews in this domain are responded to. Additionally, most responses are highly generic, as they are copy-pasted or based on templates, similar to example 1b, which does not reveal which piece of clothing or shop was reviewed. Given these premises, their goal was to automatically generate review responses that are more customised towards the customers' opinions and the products purchased. This can be seen in example 1c which mentions the store name and the material of the product. It also addresses the issue mentioned in the customer review in 1a.

(1) a. **Review:** The quality is not very good and the sleeves will pill. It looks

like the fabric is bad.

b. **Copy-pasted response:** Thank you for your support and feedback. We will continue to work hard to provide you with better service and look forward to your next visit!

c. **Specific response:** Dear customer, thank you for choosing NAISHITU flagship store. Due to the composition of polyester, slight pilling is a normal phenomenon. It is recommended that you use fabric shaver to deal with. Looking forward to your next visit. [Zhao et al., 2019]

Similar to Gao et al. [2019a], they included review-external product information in the generation process in the form of a factual table which was considered during decoding using a custom attention mechanism.

Kew et al. [2020] adapted the model architecture proposed by Gao et al. [2019a] and fed similar external information to it in order to generate responses for reviews in the hospitality domain. The data set used in this master thesis (cf. section 3.2) as well as the data set used in the study of Kew et al. [2020] are subsets of a larger collection of TripAdvisor review-response pairs that was compiled within the scope of the ReAdvisor project[1]. In contrast to this thesis' data selection, the data set used by Kew et al. [2020] also contained hotel reviews aside from restaurant reviews as well as German reviews. The experiments showed that the extended seq2seq model proposed by Gao et al. [2019a] did not outperform a basic attentional seq2seq model [Bahdanau et al., 2015; Sutskever et al., 2014] when using review-response pairs from the hospitality domain. This is in contrast to what Gao et al. [2019a] found, i.e. that their extended seq2seq model using external review information outperformed the baseline seq2seq model. Kew et al. [2020] suspected that the considerably longer average response length and the greater inter-textual variance in the restaurant and hotel reviews compared to the app reviews were responsible for the poorer performance in their adaptation.

Katsiuba et al. [2022] contributed to the research on review response generation for hospitality businesses with an analysis of how semi-automated review response generation can be implemented in different business models. Assuming that businesses outsource customer feedback management to specialised external companies, they discussed advantages and difficulties of different review generation processes including the reviewed businesses and human representatives of the feedback management companies. They found that the best solution depended on the quality of the automatically generated responses, as poorer responses need more human

[1] `https://www.cl.uzh.ch/en/texttechnologies/research/machine-learning/Response-Generation.html` (last accessed 05 October 2022).

post-processing. In a more technical part of their study, Katsiuba et al. [2022] used $\text{BART}_{\text{BASE}}$ [Lewis et al., 2020], a pre-trained Transformer-based encoder-decoder, for tackling review response generation in the hospitality domain. To guide the model towards generating responses that are adequate for either restaurants or hotels and the respective review rating, Katsiuba et al. [2022] used discrete prompting. This means they prepended discrete tokens to the review that informed on domain and review rating before inputting the review to the response generation model. Despite this additional information, a major weakness of their system was that it tended to generate universal responses that were not specific to the reviews which they responded to. Furthermore, the model sometimes "hallucinated" content that was not supported by the input review.

Kew and Volk [2022] identified the constant generation of universal responses as one of the major weaknesses of neural models when naively applied to review response generation as well. Instead of addressing aspects mentioned in the source texts, the models primarily generated highly generic answers that were legitimate for various reviews. Kew and Volk [2022] found that the hotel reviews from TripAdvisor, which they used as their data set, had a large number of many-to-one mappings, i.e. many generic responses to different reviews. This abundance of general responses caused the trained model to repeatedly generate these highly probable, and thus "safe", universal responses. As a possible countermeasure, they proposed removing review-response pairs with highly generic responses from the data set. In order to quantify a response's genericness, they tested three scoring methods which are based on either counting high-frequency words, similarity to examples of a set of generic sentences, or language model perplexity. Fine-tuning the $\text{BART}_{\text{BASE}}$ model with the filtered training sets indeed helped to reduce the genericness in the generated texts but at the cost of more unaccounted for content in the generations.

2.2 Personalised Response Generation

In this section, we first introduce the terms "persona" and "style" (2.2.1) before we delve into previous research on personalised dialogue response systems (2.2.2).

2.2.1 Terminology: Personae and Styles

A growing interest in personalisation is linked to a common problem of many conversational systems: They tend to be inconsistent [Li et al., 2016b]. This surfaces

on either the level of content[2] or style. The source of these inconsistencies lies in the training data: It is formed of utterances from different people of whom the personalities are ignored during training [Wu et al., 2021]. More recent models are trained with data sets where the utterances are attributed to different personae. The rationale behind this is that the personal traits of the underlying persona of a system should help to generate answers that are consistent with previous generated statements (e.g. on age), even without long-term memories spanning across the whole conversation history [Zhang et al., 2018].

A persona is a composition of a) identity-forming background facts and b) conversational style [Li et al., 2016b]. These two parts cannot be seen in total isolation of each other, as the identity-forming characteristics impact the conversational style. For example, the place of origin of a persona is indicative of dialectal characteristics in this persona's style. While some data sets include information on background facts such as gender or age, most encode these characteristics implicitly in the speakers' utterances (cf. chapter 3). Background facts are easily graspable, as they are concrete attributes of a person(a).

Text style, however, is a more abstract concept. Jin et al. [2022] provide and compare two definitions of style, one linguistic and one data-driven. Style in the linguistic sense is constrained by certain rules on the lexical, syntactic or discourse level, e.g. the avoidance of contractions such as "don't" in formal texts [Jin et al., 2022]. Compiling corpora that conform with such constraints is costly. This makes the linguistic definition of style inapplicable for machine learning scenarios, as especially training or fine-tuning neural models requires large amounts of data. The data-driven definition of style is more useful in such settings, since it views style as any sort of (possibly non-linguistic) characteristic that varies across text corpora [Jin et al., 2022], such as the amount of stars in restaurant reviews. In other words, data-driven styles are generally differentiated by information or metadata that is readily available without rigorous linguistic analysis. The resulting corpora, however, cannot be guaranteed to be consistent in style w.r.t. the linguistic definition.

Creating data sets for different personae using persona IDs to achieve personalised text generation fits within the data-driven definition of style, as the corpora are not distinguished by linguistic differences but metadata.

[2]For example, when asking a conversational agent for its age in two ways (e.g. "How old are you?" and "What's your age?"), it is desirable that the system answers with the same age both times.

2.2.2 Personalised Dialogue Response Generation

This section lists some of the major differences in terms of proposed architectures and methods in previous research on personalisation in dialogue response generation.

Two Kinds of Personalisation

There are two different kinds of personalisation present in the studies introduced in the following: The first and predominant kind of personalisation refers to modelling the style of a specified persona. This means that the generated text should match both with the background information of the persona (e.g. age, gender, etc.) and their conversational style [Li et al., 2016b]. However, speech style not only differs across users but also across situations. The second kind of personalisation takes into account that both the context and the addressee change a speaker's style [Holmes, 2013]. The study of Li et al. [2016b] tried to incorporate the influence of different addressees on a speaker's style by using a speaker-addressee model. The main reason why the second definition of personalisation is infrequently used in AI and NLP research is because conversations between two specific interlocutors are sparse. Investigating neural networks for specific tasks, however, requires sufficiently large amounts of training data which is not given in this scenario.

Strategies and Architectures

Text generation is a widely used method to predict a personalised response to an utterance. Not many studies investigated the alternative of retrieving possible responses from the training data [Wang et al., 2017; Zhang et al., 2018]. Smith et al. [2020] used a hybrid "retrieve-and-refine" approach where they retrieved the best-fitting response from the training set and used it to generate a new personalised response that was better adapted to the context.

Regarding the architectures that were used in these studies, both RNNs in the form of LSTMs and GRUs as well as the Transformer architecture were popular. While some used vanilla RNN-based seq2seq models with [Wang et al., 2017; Zhang et al., 2019][3] or without attention [Li et al., 2016b; Zhang et al., 2018], others extended their basic seq2seq models: Wu et al. [2021] and Zheng et al. [2019] manipulated the attention mechanism in their models. Yang et al. [2018] and Su et al. [2019] used classic RNN-based seq2seq models, but they trained their models within the dual

[3]Zhang et al. [2019] used a FFNN to generate the first token during the decoding process, the rest worked like a conventional RNN-based seq2seq model.

learning and GAN learning frameworks respectively. More recent studies employed Transformer-based models. Ma et al. [2021] trained a model that combined RNN-based and Transformer-based structures. Wang et al. [2021] used DialoGPT [Zhang et al., 2020] which is a decoder-only Transformer. Smith et al. [2020] pre-trained their own Transformer model.

Wang et al. [2021] and Smith et al. [2020] used Transformer-based models that were pre-trained on Reddit data. Others also leveraged language knowledge of pre-trained models. Li et al. [2016b] and Su et al. [2019] chose TV series characters as personae and the OpenSubtitles data set [Tiedemann, 2009] for pre-training. Another commonly used data source for pre-training material were social media platforms [Zhang et al., 2019] such as Twitter [Wang et al., 2017; Yang et al., 2018] or its Chinese equivalent Sina Weibo [Yang et al., 2018]. An advantage of leveraging social media data is the large amount of data that is readily available to learn conversational language behaviour. One noteworthy point is that none of the studies in this section used models that were pre-trained on general language data. If pre-training was applied, only data from dialogues was used for pre-training.

Indicating the Persona and Incorporating Persona Information

Another major difference between previous studies consisted in the number of models that were trained. Most studies trained a single response generation model which was capable of generating responses for different personae. For these models, it was necessary to enrich the input with information on which persona's style the response should have. A common approach to elicit responses for a desired persona was to combine a persona-specific embedding with the previous hidden state of the decoder and the last generated token [Li et al., 2016b; Su et al., 2019; Wang et al., 2021; Wu et al., 2021; Yang et al., 2018]. Another approach was to add sentences that were characteristic of a persona directly to the input prior to embedding it [Zhang et al., 2018] or to concatenate previously trained persona tokens to the input [Smith et al., 2020]. Ma et al. [2021] initialised the first hidden state of the encoder with a persona-specific vector. They also included persona information in the decoding process of the input text. Zheng et al. [2019] manipulated the attention mechanism or the output layer with explicit persona information. Some studies did not manipulate values within their models but rather adapted their outputs: Smith et al. [2020] used the non-invasive "plug-and-play" approach of Dathathri et al. [2020] which guided the generations towards being more personalised. In some studies, the researchers fine-tuned a pre-trained model for each persona [Wang et al., 2017; Zhang et al., 2019], so neither the models nor the outputs needed to be personalised.

Different ways of creating speaker-specific embeddings exist in previous studies: Li et al. [2016b] and Ma et al. [2021] trained static embeddings for each persona which were based on responses from the training data. Wu et al. [2021] produced dynamic representations by retrieving the most relevant historic utterances for the given input using an IR system and feeding them to an RNN encoder. Su et al. [2019] associated one-hot vectors with each of the characters and used the respective vector during each decoding step. Studies that utilised explicit persona information such as age or gender typically encoded each single trait as an embedding and merged the different traits' embeddings to generate a persona representation [Wu et al., 2021; Yang et al., 2018; Zheng et al., 2019]. Wang et al. [2021] used tensor factorisation to generate embeddings for input-persona pairs.

2.3 Evaluation of Personalisation

This section lists some of the most common metrics found in works on personalisation as well as those specially developed for quantifying the degree of personalisation.

2.3.1 Reference-Based Metrics

One of the most common metrics to report in response generation studies is **BLEU** [Li et al., 2016b; Ma et al., 2021; Su et al., 2019; Wang et al., 2021; Wu et al., 2021; Yang et al., 2018], a metric first developed for checking generated translations against human reference translations [Papineni et al., 2002]. To do so, it measures the word n-gram precision between a generated text and its reference(s) while punishing short and repetitive generations [Papineni et al., 2002]. In the task of response generation, the generated response is compared with one or many ground truth responses.

Another metric that measures n-gram overlap between generated and reference text is **ROUGE** [Lin, 2004], a metric that is predominantly used in the evaluation of automatic summarisation. In contrast to BLEU, ROUGE measures the recall between the generated and the gold response. The idea behind this is to measure how much the generated text has captured from the ground truth. ROUGE exists in several variants, but the only form we use is ROUGE-1 where we measure the overlap of unigrams. Kew et al. [2020] and Zhao et al. [2019] used ROUGE for review response generation evaluation, but it is less commonly reported in dialogue modelling involving personalisation [Ma et al., 2021].

Wang et al. [2021] used **F1** in their personalisation study. F1 calculates the harmonic

mean of unigram precision and recall.

Another variation of a reference-based metric is **chrF** which operates on the character level, not on the token level, and which balances recall and precision [Popović, 2015]. Kew and Volk [2022] saw an advantage of chrF compared to other token-level metrics when working with text from the internet that contains spelling errors.

Finally, it is possible to compare a generated text with a reference by **embedding** both and calculating the similarity between the vectors. Different ways of embedding the sentences and calculating similarities are possible: Ma et al. [2021] used bag-of-word embeddings and calculated average, greedy and extrema similarity. Zhang et al. [2019] used simple count vectors and calculated cosine similarity.

2.3.2 Reference-Free Metrics

One of the most frequently used metrics in the context of personalised text generation is **perplexity** [Li et al., 2016b; Smith et al., 2020; Wang et al., 2021; Wu et al., 2021; Yang et al., 2018; Zhang et al., 2018; Zheng et al., 2019] which is calculated with a pre-trained language model. While it does not reveal whether a system's generated texts display a large degree of personalisation, it is a common measure to evaluate the outputs' fluency and grammaticality [Jin et al., 2022]. Some weaknesses of perplexity include that shorter sentences score better than longer ones and that frequent words result in better perplexity than less common terms [Jin et al., 2022]. Furthermore, perplexity is highly dependent on the language model's architecture and training data as well as the model's training procedure, so using different models to calculate perplexity results in different scores [Jin et al., 2022].

Other reference-free metrics measure the diversity of the generated texts. **Distinct-N** as used by Li et al. [2016a] counts the distinct number of unigrams for Distinct-1 (and the number of bigrams for Distinct-2) *across* all generated responses and divides it by the total number of unigrams (or bigrams) in all responses. Kew and Volk [2022] applied this metric to the n-grams *within* a single response: They counted the unique unigrams (and bigrams) within a response and scaled this value by the total number of unigrams (or bigrams) in the response. For the final score, they calculated the average over all responses' Distinct-1 (or -2) scores. The first method is useful to gauge the overall inter-textual diversity of the generated texts. Inter-textual diversity is of interest because response generation systems tend to produce highly generic answers that fit various inputs [Kew and Volk, 2022] which is an undesired behaviour. A higher score in inter-textual diversity indicates that the model generates responses that are either more tailored towards the inputs or more

incorrect — introspection is necessary to determine the source of increased diversity. The second implementation of Kew and Volk [2022] is helpful for estimating the generated texts' intra-textual diversity. Low intra-textual diversity is an indicator for repetitive loops in which neural models tend to get stuck [Holtzman et al., 2019].

Self-BLEU [Zhu et al., 2018], another inter-textual diversity metric, adapts the BLEU metric. However, instead of checking each generated response against a ground truth response, the comparison is made with the other generated responses. The average BLEU score of all generated responses is the Self-BLEU score. A high Self-BLEU score implies that there are many n-gram repetitions across the generated responses which indicates that the responses are less diverse.

Kew and Volk [2022] reported two further simple metrics: The total number of **unique words** is indicative of the lexical range of the model. The **average length** of a generated response provides an approximate idea of its specificity, as shorter texts cannot adequately address topics raised in the review.

2.3.3 Personalisation Metrics

In this section, we describe all metrics of which we know that are designed for measuring the degree of personalisation. These metrics have in common that they use the personae's training set responses in some way: While some metrics compare the generated responses directly to the training set responses, others are based on models which are trained with the training set responses (such as the language models used for calculating Per-Hits@k as can be seen below). The reason why using the entire training set is preferred over single or small numbers of references is that the larger amount of data in the training sets potentially captures more of a persona's characteristics. This is relevant considering that the number of possible responses for any input is infinite, i.e. response generation is an open-ended task.

Persona-R/-P/-F1, for example, measures the unigram recall, precision or F1 score (without considering stop words) between a generated text and the text data available from the persona's training set [Lv et al., 2020; Ma et al., 2021].

P-Cover [Ma et al., 2021; Song et al., 2019], which is defined in equations 2.1 to 2.3, goes a step further by giving more weight to less frequent words in the unigram overlap. Specifically, the averaged inverse term frequency (ITF)[4] of all tokens from

[4]Song et al. [2019] called this the "inverse document frequency", but this is a) likely to be confused with the inverse document frequency from the TF-IDF measure, and b) not what is expressed in formula 2.1. Miyazaki et al. [2021] noted this error as well, and therefore, we renamed this variable to "inverse term frequency" (ITF).

the overlap set W of words occurring in the generated text $\hat{y}_i$ and the persona text p_j is calculated for each of the M persona texts. Of all these weighted overlap scores, the largest one is considered for each generated text, the idea of this being that this is the persona text most similar to the generated text. The overall P-Cover is averaged across all N generated texts for the respective persona.

$$ITF_i = \frac{1}{(1 + log(1 + TF_i))} \quad [5] \tag{2.1}$$

$$S(\hat{y}_i, p_j) = \frac{\sum_{w_k \in W}(ITF_k)}{|W|} \tag{2.2}$$

$$C_{per} = \frac{\sum_i^N max_{j \in [1,M]} S(\hat{y}_i, p_j)}{N} \tag{2.3}$$

Another consequence of the open-endedness of response generation is that not all elements in a pool of references express the same semantics. Therefore, Xu et al. [2018] noted that calculating the BLEU score over all references is not reasonable for response generation. In their proposed metric **MaxBLEU**, they grouped semantically similar references and only considered the largest BLEU score between the generated response and any of the reference groups.

Su et al. [2019] adapted this idea and transformed it to an accuracy measure[6] which we call **MaxBLEU accuracy**. This metric calculates the MaxBLEU score between a generated response and each persona set individually. Specifically, it calculates the BLEU score between the generated response and each sample from the persona's training set separately[7] to only consider the largest BLEU score as a persona set's MaxBLEU score. It then calculates the percentage of generated responses that have the highest MaxBLEU score with the persona towards which they were personalised.

Wang et al. [2021] developed a metric that includes perplexity to evaluate the degree of personalisation, i.e. **Per-Hits@k**. Given that there are N personae and M_i

[5] TF is "computed from the [frequency] index via Zipf's law": $TF_i = 1e6 * \frac{1}{(idx_i^{1.07})}$ [Zhang et al., 2018]. Zipf's law states that a term's frequency is inversely proportional to its frequency rank [Piantadosi, 2014]. The frequency rank is obtained by ranking terms by their frequency, i.e. $idx = 1$ for the most common term, $idx = 2$ for the term with second-highest frequency, etc.

[6] The conversion to an accuracy measure only becomes evident from looking at their source code: `https://github.com/adelaidehsu/Personalized-Dialogue-Response-Generation-from-Monologues/blob/85f1b9596cfb6abfe5a0ab803ff860203579de15/main.py#L831` (last accessed 01 September 2022).

[7] While Xu et al. [2018] created groups of semantically similar references, Su et al. [2019] calculated the BLEU score between each persona set sample and the generated response separately.

responses generated for persona i, Per-Hits@k requires a user to train N language models. Each response that is generated for a persona is tested on all personae's language models which results in N perplexity values. The perplexity of the persona's own language model is then ranked over all other perplexities, and the final value of Per-Hits@k is calculated as in equation 2.4.

$$Per\text{-}Hits@k = \frac{1}{\sum_{i=1}^{N} M_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} 1_{x \leq k}(rank(ppl_{i,j}^{i})), \tag{2.4}$$

where $ppl_{i,j}^{n}$ is the perplexity of persona i's j-th response on persona n's language model. Thus, Per-Hits@k quantifies the probability that the generated responses' perplexities rank in top-k with the respective personae's language models [Wang et al., 2021].

PTSal (Persona Term Salience) [Miyazaki et al., 2021] adopts the idea that some words are more important when evaluating personalisation than others, similarly to P-Cover [Song et al., 2019]. The words are not weighted by frequency as for P-Cover [Ma et al., 2021; Song et al., 2019] but by the degree to which they are characteristic of a persona. This idea is an adaptation of TF-IDF which is used to gauge the importance of a term for a document in information retrieval. Like TF-IDF, PTSal is highest when a term occurs frequently in a small number of persona training sets and lowest when a term occurs in all persona training sets [Manning et al., 2008]. The PTSal of each term and persona is first calculated with a persona's training set as in equations 2.5 and stored in a table. The overall PTSal score of a response is the average PTSal of all its tokens.

$$\begin{aligned} PTSal(t,p) &= UttFreq(t,p) * SpkrRarity(t) \\ UttFreq(t,p) &= \tfrac{n(t,p)}{m(p)} \qquad SpkrRarity(t) = log\tfrac{|P|}{s(t)}, \end{aligned} \tag{2.5}$$

where $n(t,p)$ is the number of training set responses of persona p with term t, $m(p)$ is the number of training set responses of persona p, $|P|$ is the number of all personae, and $s(t)$ is the number of personae that used term t. Analogously to TF-IDF, $UttFreq$ describes how often a term is used by a persona, whereas $SpkrRarity$ measures how characteristic a word is of a persona style. Miyazaki et al. [2021] found that their metric PTSal correlated weakly to moderately with human judgement on how characteristic a text is for a persona.

Smith et al. [2020], Su et al. [2019], and Zheng et al. [2019] used a **classifier**-based

approach to personalisation evaluation. The first two trained a response classifier to predict the persona style. Zheng et al. [2019] trained a classifier for each persona trait (e.g. gender, location, etc.) for which the text generation could be conditioned. The classification accuracy informed about how much the generated responses conformed to the respective personae's or persona traits' style. A study on text attribute transfer by Li et al. [2018] showed that such a classifier's accuracy score correlated variably on different data sets with human judgement.

In section 4.2.3 we list the personalisation metrics we use in this thesis. Section 6.4 discusses some of the pitfalls of the introduced metrics.

2.4 Leveraging Pre-Trained Models

With the ascent of large language models such as BERT [Devlin et al., 2019] or GPT [Radford et al., 2018] which are pre-trained on enormous amounts of text, the main paradigm of model creation for downstream tasks such as response generation is fine-tuning a pre-trained language model. While fine-tuning often results in a strong performance in the respective task, one copy of the new fine-tuned model needs to be stored individually for each downstream task. This occupies prohibitively large amounts of memory. Another disadvantage of fine-tuning occurs in low-data scenarios: The weights of the pre-trained model need to be sufficiently close to the optimal weights of the downstream task for fine-tuning to be successful [Phang et al., 2018]. When updating weights of a complex model with only a small number of data points, we additionally run risk of overfitting the model to the data samples [Géron, 2019]. Another risk of fine-tuning is catastrophic forgetting where the pre-trained model unlearns knowledge acquired during the pre-training tasks while being fine-tuned [Kirkpatrick et al., 2017].

An alternative that reduces the memory load of re-tuned parameters and minimises the risk of overfitting as well as catastrophic forgetting is "light-weight fine-tuning" where only a subset of the large language model's weights or some additional weights are tuned and stored [Li and Liang, 2021]. Since these methods preserve the majority or all of the pre-trained model's weights, they are likely to generalise well to examples unseen during fine-tuning, as their original weights have been pre-trained on large amounts of data from various domains. In other words, these lightweight fine-tuning methods preserve a model's inductive bias better than conventional fine-tuning.

2.4.1 Prefix-Tuning

Prefix-tuning [Li and Liang, 2021] is a lightweight fine-tuning approach where a small number of parameters is tuned for a downstream task. These parameters are prefixed to the pre-trained model to solve the task. Figure 2 shows a schematic view of the difference between the tunable weights in fine-tuning and prefix-tuning: Fine-tuning updates and stores all parameters of the pre-trained model for each individual downstream task. Meanwhile, prefix-tuning requires us to optimise only a handful of additional weights for each downstream task, while the pre-trained model remains unchanged. Thus, for prefix-tuning, only the prefixes and one version of the pre-trained model need to be stored.

There are some advantages of prefix-tuning over fine-tuning [Li and Liang, 2021]:

1. Only one prefix has to be stored per downstream task rather than a whole fine-tuned model. This makes prefix-tuning a highly memory-efficient approach.
2. Using prefixes allows one to assemble queries from different users in the same batch to process these efficiently on the same pre-trained model. Batching works because the suitable task-specific prefix can be modularly prepended to each query without changing the model.
3. Prefix-tuning outperforms fine-tuning in the low-data setting and in extrapolating to topics unseen during task adaptation.[8] Prefix-tuning is almost on par with fine-tuning when using sufficiently large data sets.

Prefix-tuning prepends freely tunable weights to the activations of all model layers. The weights on the embedding layer are not restricted to correspond to real tokens which increases their expressivity [Li and Liang, 2021]. The prefixed weights on the subsequent layers have the advantage that they can influence the calculations in the deeper layers of the network more directly than the prefixed weights on the embedding layer [Li and Liang, 2021]. If the pre-trained model is an encoder-decoder, separately tunable prefixes are added to both the encoder and the decoder [Li and Liang, 2021]. The intuition behind this is that the encoder prefix influences what the model extracts from the input, whereas the decoder prefix affects the output generation by adapting the distribution of the next token [Li and Liang, 2021]. Generally, autoregressive language models calculate intermediate activations as in equation 2.6 [Bahdanau et al., 2015; Li and Liang, 2021].

[8] Li and Liang [2021] observed this for the table-to-text and summarisation tasks. Although this is promising, it is unclear yet whether this observations extrapolates to other tasks.

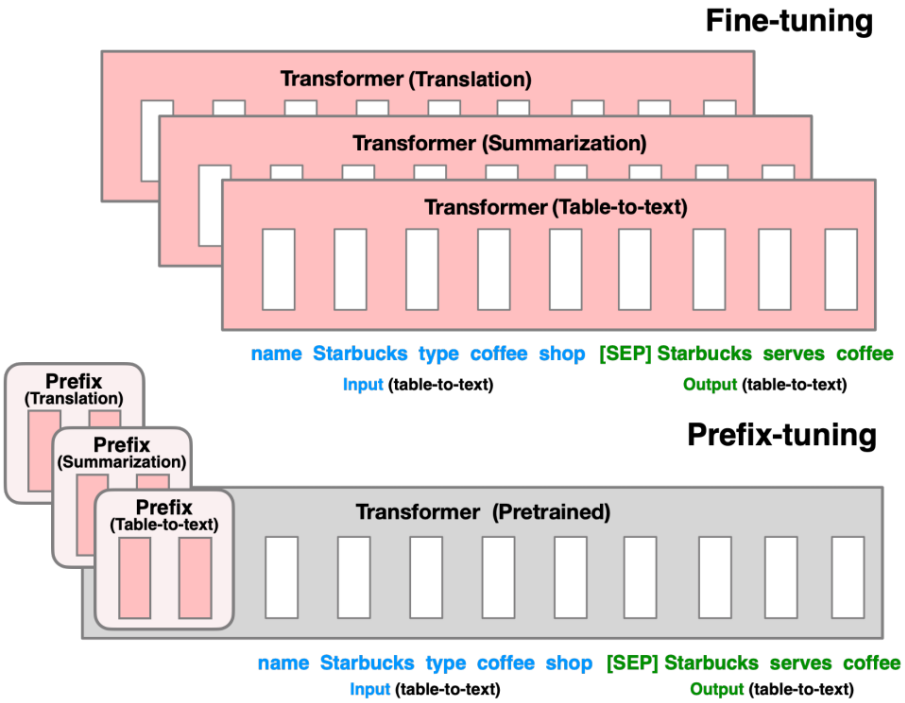


Figure 2: Comparison between fine-tuning and prefix-tuning for adapting a pre-trained model to a downstream task (illustration copied from Li and Liang [2021]). Red colouring indicates that the weights are tuned and stored.

$$h_i = LM(y_{i-1}, h_{<i}), \tag{2.6}$$

where h_i is the activation vector at time step i and y_i is the generated token at time step i during inference or the ground truth token at position i during training[9] [Li and Liang, 2021]. Using prefix-tuning, the activation vector at the prefix indices consists of the prefix values as can be seen in equation 2.7 [Li and Liang, 2021].

$$h_i = \begin{cases} P_\theta[i, :], & \text{if } i \in P_{idx}, \\ LM(y_{i-1}, h_{<i}), & \text{otherwise}, \end{cases} \tag{2.7}$$

where P_θ is the tunable prefix matrix $\in \mathbb{R}^{|P_{idx}| \times dim(h_i)}$ and P_{idx} is the set of prefix indices[10]. In other words, when $i \in P_{idx}$, h_i copies from the prefix matrix. When

[9] Inputting the last time step's ground truth token at every time step to the model rather than the previously generated token is called teacher forcing [Goodfellow et al., 2016]. This ensures that the model's predictions during training are always based on the ground truth history.

[10] The size of the prefix is a hyperparameter set by the user. The number of prefix indices equals the prefix size.

$i \notin P_{idx}$, h_i is computed by the language model, but h_i is still affected by the prefix weights on the left [Li and Liang, 2021]. The prefix-tuning illustration in figure 2 shows that the prefix weights are copied for the activation vectors at the positions in P_{idx}, while the activations from the other positions are calculated using the model.

3 Data for Personalised Text Generation

To apply prefix-tuning in the context of personalised restaurant review response generation, we create a data set that has sufficient data from a suitable number of restaurants. In section 3.1, we first review some studies in which custom personalisation data sets were compiled. Our goal is to get a better understanding of which considerations previous researchers made when creating personalisation data sets. Section 3.2 describes how we compile our own data set of review-response pairs. It also contains some quantitative descriptions of the data set and lists the pre-processing steps.

3.1 Personalisation Data Sets in Previous Studies

The following data sets were created in studies on personalised text generation. This means that each data point in these data sets is attributed to a persona ID.

We have to keep some differences between the following data sets and the one we plan to compile in mind:

1. None of the following data sets includes review-response pairs. They rather focus on conversational data mostly from social media that does not follow a similarly rigid structure as it is characteristic of hospitality business reviews and responses [Pantelidis, 2010; Zhang and Vásquez, 2014].

2. The models trained in these studies are not limited to a single domain. This is a consequence of using conversational training data that raises many different subjects [Zhang et al., 2018]. In contrast, we want to solve the task of review response generation which is limited to a few topics.

3. A characteristic of review response generation that sets it apart from chit-chatting is its single-turn nature. This is predetermined by the entry interface of typical reviewing platforms that limits interaction to single turns.

4. We consider all responses of a business to originate from the same persona, although we do not know how many people wrote responses for each individual business. In contrast, the data sets from the previous studies only consider text written by a single person per persona.

Despite these differences, we consider the data set sizes and persona choices made in previous works to create our data set for personalised review response generation.

3.1.1 Data Sets With Explicit Persona Information

Some data sets not only attribute utterances to a persona ID, but they also enrich these IDs with background information of these personae. This kind of data set is used in the following studies: One of the best-known data sets for personalised response generation is PERSONA-CHAT [Zhang et al., 2018]. PERSONA-CHAT collects a total of 1,155 fictive personae that are based on their personal interests. All of the personae are described by at least five profile sentences similar to the ones in example 2.

(2) I am a vegetarian. I like swimming. My father used to work for Ford. My favorite band is Maroon5. I got a new job last month, which is about advertising design. [Zhang et al., 2018, 2206]

To compile this data set, crowdworkers were randomly assigned a persona and asked to chat with a fellow crowdworker to produce dialogues between two personae. This resulted in a total of roughly 165,000 utterances.

The `PersonalDialog` data set [Zheng et al., 2019] is larger than the PERSONA-CHAT data set with 56.25 million utterances from 8.47 million speakers. These utterances are part of multi-turn dialogues taken from Weibo, a Chinese social media platform. User traits are publicly available on the users' profiles and include gender, age, location, interest tags, and self description. Zheng et al. [2019] relied on these personality traits to generate personalised responses that were typical of the conversational behaviour of the respective demographic groups.

Yang et al. [2018] also used data from Weibo for a Chinese model and data from Twitter for an English model. They first used conversation data without explicit user information to pre-train their conversational model. After pre-training, they used conversation data augmented with user information to fine-tune the model for personalised response generation.

PER-CHAT [Wu et al., 2021] is compiled from Reddit's subthread r/AskReddit where users' open-domain questions are answered by other users. This results in over 1.5 million question-answer pairs from 300,000 users. The data set is augmented with user IDs, the Reddit comment histories of all users, and user profile attributes similar to the ones of `PersonalDialog` [Zheng et al., 2019]. Wang et al. [2021] used a subset of PER-CHAT for tackling personalised response generation.

3.1.2 Data Sets Without Explicit Persona Information

Acquiring explicit personal information is not always possible due to privacy restrictions or lack of financial resources [Ma et al., 2021]. Therefore, other approaches to personalisation only use user utterance histories to customise conversational agents. The only information given in this scenario is text that is linked to a user ID.

Mazaré et al. [2018] noted that PERSONA-CHAT [Zhang et al., 2018] is completely artificial, i.e. its utterances were not produced in real-life circumstances. As a consequence, they raised the doubt that such artificial data may not be suitable to represent authentic user-bot interactions. To provide a real-life data alternative, they built a data set with data from Reddit consisting of over 700 million dialogues from over five million users. All context-response pairs (the context is the utterance that is responded to) are accompanied by sets of sentences that characterise the responding person (e.g. "I like sport" or "I work a lot"). The sentences are collected from the users' comment histories based on simple rules.

Ma et al. [2021] worked with data from Weibo and Reddit. Both the Chinese and the English data sets are smaller than the one of Mazaré et al. [2018]. Li et al. [2016b] worked with social media data as well. They took message-response pairs of 74,000 users from Twitter to create speaker embeddings for personalised text generation.

Smith et al. [2020] controlled response generation only for conversational style. Their data set Image-Chat [Shuster et al., 2020] comprises 217 styles (like "curious" or "casual") and approximately 400,000 utterances. For each dialogue, two crowdworkers were asked to chat with each other about an image in the style that was allocated to them. This data set is therefore as artificial as PERSONA-CHAT, as it was also created by crowdworkers.

Su et al. [2019] trained a conversational agent to produce personalised answers for the main characters of the TV series *Friends* and *The Big Bang Theory*. To this end, they used part of the OpenSubtitles data set [Tiedemann, 2009] to pre-train the model and provide it with general knowledge on natural conversations. For per-

sonalisation they collected roughly 170,000 lines of monologic speech of the series' characters. All in all, Su et al. [2019] only required a corpus with general conversations and a corpus of monologues for personalisation rather than dialogues with personal information.

Wang et al. [2017] took a similar approach to Su et al. [2019]: They first used a larger open-domain dialogue corpus (the Twitter Conversation Triple Data set by Sordoni et al. [2015]) for instilling general conversational knowledge to the model. They then used smaller corpora containing monologues for all personae for which they wanted to generate personalised responses. The monologues belonged to different genres and were spoken by real as well as fictional characters, such as the politicians Hillary Clinton, Donald Trump, John F. Kennedy, and Richard Nixon. They also used texts from singer-songwriters, stand-up comedians, and Star Wars characters.

Zhang et al. [2019] also pre-trained a model on general conversational data from Chinese online forums before adapting the model to the style of five volunteers. Using 2,000 single-turn conversation pairs from each of them, they created five personalised models.

3.1.3 Demographics of Personalisation Data Sets

The number of personae in the aforementioned studies is highly varying from using five up to 8.47 million personae as can be seen on table 1. There are some patterns that emerge from this table: Generally, data sets with a large number (i.e. several thousands up to millions) of personae had fewer personalised utterances than data sets with maximally a few hundred personae. The data sets with the fewest personalised utterances [Wu et al., 2021; Zheng et al., 2019] compensated the personal data sparseness by using explicit user information. In contrast, none of the studies with few personae except for Yang et al. [2018] included explicit persona information.

Another difference between studies that personalise for thousands of personae and smaller-scale personalisation studies is their data source: Those with few personae had more diverse data sources. The studies with the largest numbers of personae typically worked with data crawled from social media platforms. Some of the smaller-scale studies used data that was exclusively created by crowdworkers or volunteers to train personalised response systems [Smith et al., 2020; Zhang et al., 2018, 2019].

The last difference between studies with few and many personae is that the former exclusively used pre-trained models, whereas the latter typically trained their models from scratch. This is plausible, as the overall number of utterances to train on is

Study	Data source	Personae/ styles	Avg. utterances per persona[1]	explicit persona info	pre-training
Smith et al. [2020]	crowdsourced	217	1,843	✗	✓
Su et al. [2019]	TV series	13	12,947	✗	✓
Wang et al. [2017]	political speeches, music, movies, shows	7	∼6,000	✗	✓
Yang et al. [2018]	Twitter	178	631	✓	✓
Zhang et al. [2019]	volunteers	5	2,000	✗	✓
Wang et al. [2021]	Reddit	4,724	42	✓	✓
Zhang et al. [2018]	crowdsourced	1,155	142	✓	✗
Li et al. [2016b]	Twitter	74,003	92	✗	✓
Ma et al. [2021]	Reddit	300,000	29	✗	✗
Mazaré et al. [2018]	Reddit	5M	140	✗	✗
Wu et al. [2021]	Reddit	301,243	5	✓	✗
Zheng et al. [2019]	Weibo	8.47M	7	✓	✗

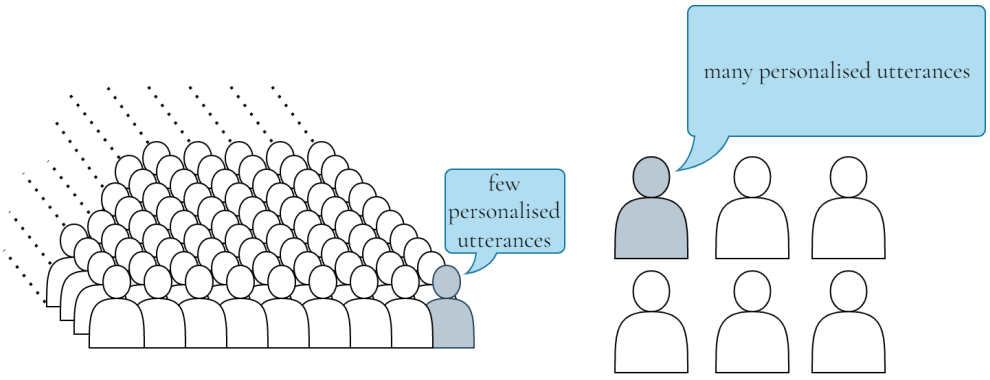
Table 1: Overview of personalised data sets used in previous research. Green rows refer to studies with few personae, orange rows to studies in mid-range w.r.t. number of personae, blue rows to studies with many personae.

much larger in those studies with hundred thousands of personae than in those with some tens of personae. Figure 3 summarises these findings.

3.2 The TripAdvisor Data Set for Personalised Review Response Generation

The previous sections have shown that data sets for personalised response generation may include only a handful of personae up to several millions. Generally, the more personae a data set comprises, the fewer utterances are available per persona. A persona in our case is a restaurant. The distinction between personae that are users in online forums, fictional or real people, and hospitality businesses such as restaurants is related to a difference in the function of the response generation systems: Systems

[1]For the studies of Mazaré et al. [2018] and Zhang et al. [2018], the number of utterances does not include the number of persona sentences describing a persona. For the study of Wu et al. [2021], the utterance count only includes personalised question-answer pairs but not the comment history of the users since the size of the comment history cannot be deduced from the paper.



Personalisation for many personae	Personalisation for few personae
personalisation for a few hundred thousand up to several millions of personae	personalisation for only a handful of personae (less than ten up to a few hundred)
personalisation data from social media	personalisation data from individuals, TV series, politicians' speeches, crowdsourcing, etc.
sometimes with explicit persona information such as age, gender, location, etc.	typically no explicit persona information
often less than 100 utterances per persona	a few hundred up to a few thousand utterances per persona
conversational model often trained from scratch	conversational model typically pre-trained for chit-chatting on large data sets

Figure 3: Personalising conversational models for many personae vs. few personae

trained to answer like the first kind of personae can be considered as open-domain chit-chat systems [Zhang et al., 2018]. Systems adapted to responses of hospitality companies, on the other hand, are oriented towards providing reputation-enhancing responses to customer reviews. The scope of topics that such systems need to address is limited to a few domain-related ones, such as food, service, special events, etc. Since review responses are highly conventionalised, meaning that similar rhetorical moves reoccur across different responses [Zhang and Vásquez, 2014], the limited diversity in style and content is learnable by a model using a limited amount of data.

While a response generating system does not produce reviews, it has to process them to generate responses. Review and response are therefore "intertextually connected" [Zhang and Vásquez, 2014], as the latter cannot exist without the former. Pantelidis [2010] found the following preference structure for restaurant reviews w.r.t. the content: "food, service, ambience, price, menu, and decor (in that order)". Not all of these topics occur in each review, but these are the most frequent aspects rated by the reviewers. The review topics dictate the topics that the response contains.

Zhang and Vásquez [2014] found that the rhetoric moves in hotels' responses to reviews on TripAdvisor were similar across different hotels and included the expression of gratitude, apologies for sources of trouble and proofs of action (in responses to

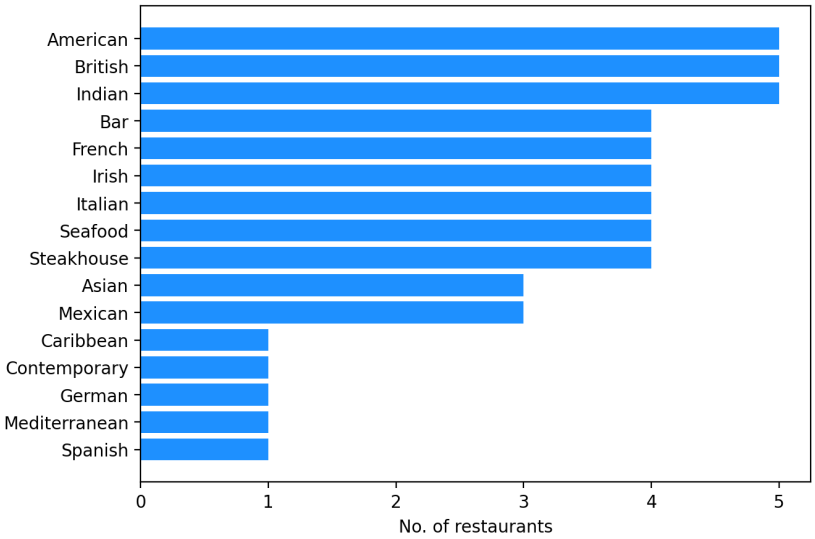


Figure 4: Cuisine types[2] of data set restaurants

reviews with negative ratings), invitations to future visits, acknowledgements of the feedback provided in the review, and references to the customer review. Although their analysis did not entail restaurant reviews, we assume that similar patterns exist within this domain, since both hotels and restaurants are part of the hospitality business. Like reviews, review responses exhibit a preferred structure where, for example, the expression of gratitude is typically located at the beginning of the response [Zhang and Vásquez, 2014]. While both reviews and responses are highly conventionalised, there are still cross-cultural [Napolitano, 2018] and idiosyncratic differences in the responding styles. For this reason, we compile a data set comprising review-response pairs with responses of 50 different restaurants from the United States, United Kingdom, Australia and Ireland (cf. figure 5). This data set is a subset of the TripAdvisor data set collected as part of the ReAdvisor project[3].

We select the review-response pairs from 50 of the top-80 restaurants with the most answered reviews. Since 24 of the top-50 restaurants are serving either British or American food, we replace those of the British and American restaurants with fewer review-response pairs with restaurants from the top-80 restaurants that specialise in other cuisines. We do so to include more diversity across the personae. Table

[2]Cuisine types according to first category in categorisation on TripAdvisor. We retyped "European" to the next, more specific category on TripAdvisor, e.g. "British". We did not apply the same retyping to the category "Asian" due to the small number of restaurants in this category.

[3]`https://www.cl.uzh.ch/en/texttechnologies/research/machine-learning/Response-Generation.html` (last accessed 05 October 2022).

6 in appendix A shows the selected restaurants, their locations, price categories, and numbers of review-response pairs. Food is the most prevalent topic in customer reviews [Pantelidis, 2010; Parikh et al., 2017], and review responses are dependent on the reviews. Thus, choosing restaurants that serve different cuisines is likely to help us select more diverse personae that have more distinctive responses. The distribution of the restaurants' cuisines is given in figure 4.

To further encourage diversity across the review-response pairs of the different personae, we include restaurants from different countries and varying price categories in the data set (cf. figure 5). Two thirds of the restaurants are categorised as "mid-price" on TripAdvisor, whereas one third is located at the higher price end. Only one restaurant belongs to the category "cheap eats", as restaurants with many review-response pairs from this price category are scarce. We assume the reason for this is that restaurants in this budget range do not have the personal or temporal resources to write thousands of responses to online reviews.

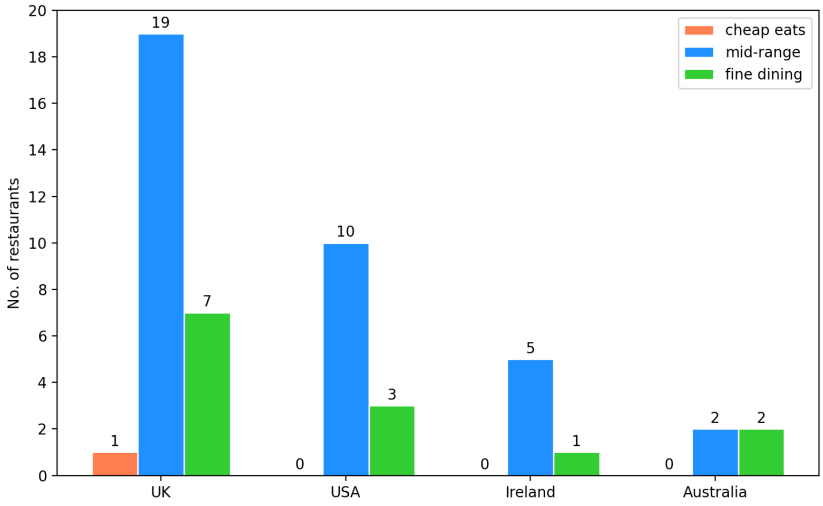


Figure 5: Locations and price categories of restaurants in our data set

With this selection of restaurants as personae, our approach differs from other work on personalisation [Ma et al., 2021; Mazaré et al., 2018; Wu et al., 2021; Zheng et al., 2019] that uses data sets with thousands up to millions of different personae but less than 1,000 utterances per persona. Having fifty restaurants as personae with an average of 2,000 review-response pairs, it is more comparable to previous work [Smith et al., 2020; Su et al., 2019; Wang et al., 2017; Zhang et al., 2019] that limits personalisation to a smaller set of five up to roughly 200 personae. In these data sets, each persona is generally linked to several thousands of utterances. Section 3.1.3 shows a more complete overview over other studies' personalisation data sets.

Our final data selection has 100,011 review-response pairs, on average 2,000 per

restaurant. Figure 8.a) shows the distribution of review-response pair counts across the data set's restaurants.

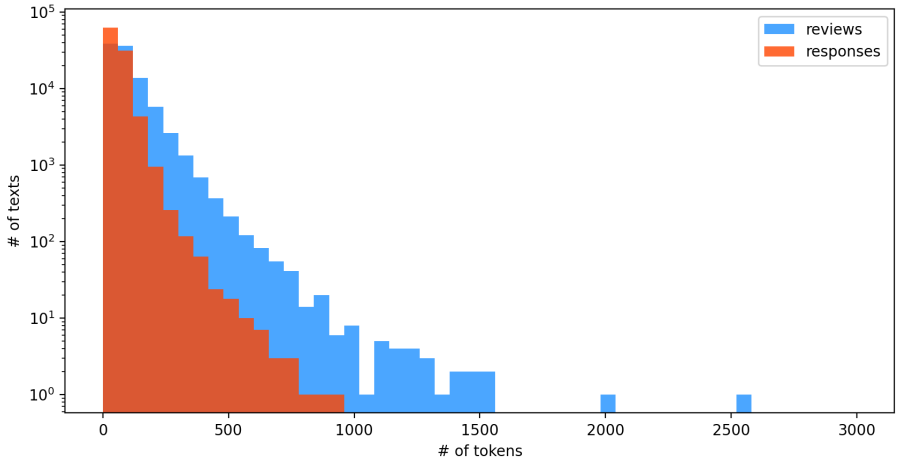


Figure 6: Distributions of review and response lengths in our data set prior to pre-processing

The mean length of reviews is 6 sentences and 98 tokens. For responses, it is 3.9 sentences and 58 tokens. Thus, responses are typically shorter than reviews. The text lengths' distributions of both reviews and responses is shown in figure 6. Texts of increasing length are decreasingly frequent up to a point where there are only some outliers. Figure 8.c) shows the quartiles of response lengths across different restaurants before and after pre-processing. The pre-processing steps, which we list in section 3.2.1, generally shorten the responses by a few tokens.

Table 2 shows the sizes of our training, validation, and test splits. We shuffle the data and apply stratified sampling to ensure that the different splits have the same percentage of samples from each restaurant as in the original data set.

Split	RR Pairs	# of sentences in reviews	# of sentences in responses
Training	80,011	480,173	301,321
Validation	10,000	59,453	37,742
Test	10,000	60,551	37,728

Table 2: Overview of training, validation, and test splits of our TripAdvisor review-response data set for personalisation. We tokenised the sentences using NLTK's[4] `sent_tokenize` function after applying all pre-processing steps.

Shuffling also ensures that the percentage of one- to five-star reviews is similar across the splits. Figure 7 shows the composition of one- to five-star reviews in

[4] `https://www.nltk.org/index.html` (last accessed 08 November 2022).

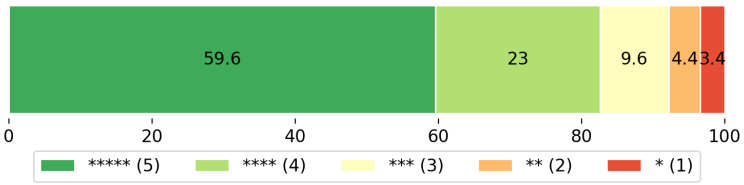


Figure 7: Percentages of review ratings in our data set

the complete data set, and figure 8.b) illustrates the distribution of average ratings across the data set's restaurants. Similar to what Pantelidis [2010] found in his analysis of online restaurant reviews, the major part of our reviews is positive with four or five stars. The fewer-stars reviews become less frequent with a decreasing number of stars. Section 6.2 analyses the consequences of this underrepresentation of negative reviews for the quality on the generated responses.

Regarding the uniqueness of the responses, figure 8.d) suggests that around 75% of the restaurants have at least 9 out of 10 unique responses in their data sets. The restaurants from the lower whisker have around two thirds up to 90% of unique responses. Three outlier restaurants have fewer unique responses. After pre-processing, the percentage of unique responses generally becomes lower, and the number of the bottom outliers increases by two. This indicates that some responses are only unique by means of small differences that are erased during pre-processing, such as addressing different users in the greetings.

Figure 8.e) shows the average mentions of the restaurants' names in their responses. Some restaurants almost never mention their names, while some mention them more than once per response.

3.2.1 Pre-Processing

We pre-processed the data in the following way: First, we masked all greetings and salutations in the responses, as they contain information that our text generation system does not have access to, e.g. the user name of the reviewer or the name of the respondent. Additionally, greetings and salutations are highly standardised and thus can be easily inserted into the responses after they are generated. We used an existing sequence labelling model created in the context of the ReAdvisor project[5] to

[5] `https://www.cl.uzh.ch/en/texttechnologies/research/machine-learning/Response-Generation.html` (last accessed 05 October 2022).

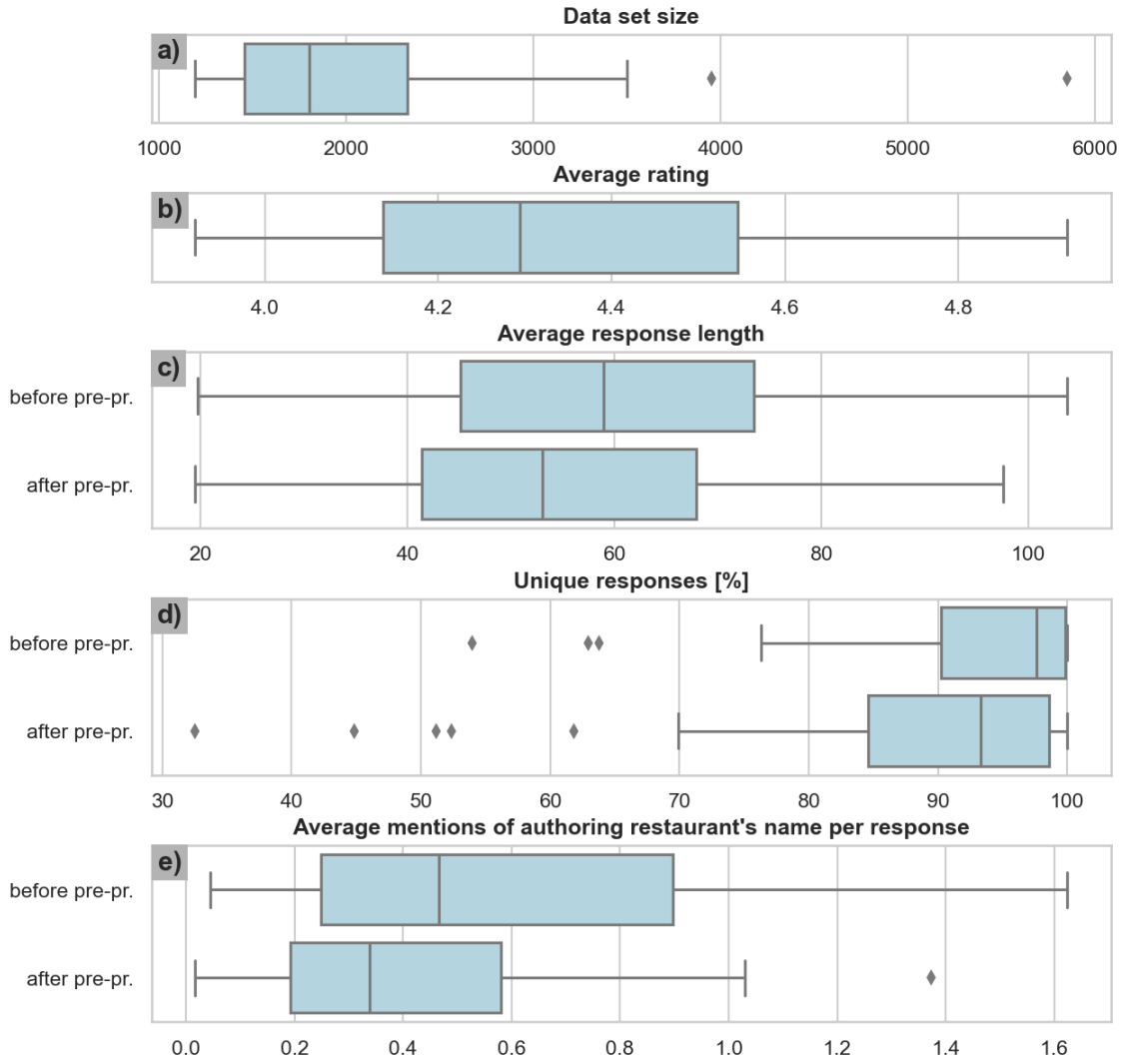


Figure 8: Distributions of some quantitative descriptions of our data set across restaurants. Plots c), d), and e) differentiate distributions prior to and after pre-processing. All distributions except for a) are measured only on the training set, a) includes the whole data set.

mask greetings and salutations in review responses. The neural network uses Flair embeddings [Akbik et al., 2018]. It consists of a bi-directional LSTM which feeds its hidden representation at each time step to a linear layer to tag it with a greeting, a salutation, or a response body tag. The model was trained on 800 review-response pairs from the hospitality domain in German and English. 100 review-response pairs each were used in the validation and test splits. On the test set, a macro accuracy of 90% was achieved for the labels ⟨GREETING⟩ and ⟨SALUTATION⟩. As for further pre-processing steps, we replaced all email addresses and URLs with special tokens

using the NLP framework spaCy[6].

Figure 8.c) shows that pre-processing generally shortens the responses. The reason for this is that we replace multi-token greetings and salutations with single-token tags. Figure 8.d) indicates that removing greetings and salutations lowers the percentage of unique responses for a few restaurants. This suggests that some of the responses of some restaurants are identical except for their greetings and/or salutations. In other words, copy-pasting responses while only adapting greetings and salutations is an existing practice among response writers. Figure 8.e) suggests that pre-processing has an impact on the number of restaurant name mentions. We assume that many restaurants mention their names in the salutations. By masking them, many of the restaurant mentions are removed.

[6] `https://spacy.io/` (last accessed 29 September 2022).

4 Review Response Generation via Prefix-Tuning

In this chapter, we specify how we adapt prefix-tuning [Li and Liang, 2021] for our task. In section 4.1, we explain how we use prefixes for personalised review response generation. In section 4.2, we describe our experiments and training procedure and list the evaluation metrics. In section 4.3, we show the results of some preliminary experiments which we conducted to determine the optimal prefix lengths.

4.1 Personalisation Using Prefixes

Figure 9 shows which weights a user has to tune and store for fine-tuning and prefix-tuning in the case of personalised response generation. Fine-tuning implies tuning and storing all weights for each fine-tuned model for each restaurant. Prefix-tuning requires only one copy of the original large language model and stores the considerably smaller prefix weight matrices for each restaurant.

How does our approach to personalisation differ from the previous studies described in section 2.2.2? Our take on personalisation corresponds to the most common interpretation where personalisation refers to the adaptation of a persona's style, in our case the style of a restaurant. Unlike most studies and same as Smith et al. [2020] and Wang et al. [2021], we do not use an RNN architecture but a Transformer-based seq2seq model. More specifically, we use $\text{BART}_{\text{LARGE}}$ [Lewis et al., 2020]. As we include only few personae (compared to other data sets in section 3.1) and thus relatively little training material, we profit from the pre-training of BART, similar to others who also pre-train their models and work with a comparable amount of personalisation data [Smith et al., 2020; Su et al., 2019; Wang et al., 2017; Yang et al., 2018; Zhang et al., 2019]. In our approach to personalisation, we use a single base model (instead of tuning as many models as personae), similar to what the majority of the studies do. Like others, we indicate the respective persona to the model by adding a persona representation to the text generation process. A main

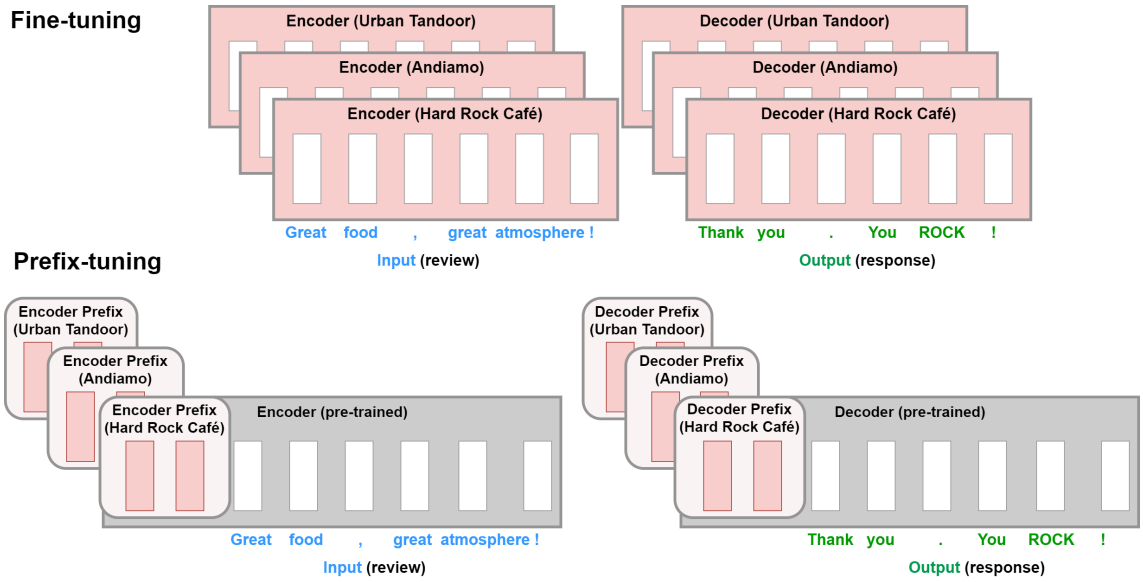


Figure 9: Schematic drawing comparing fine-tuning with prefix-tuning for personalised review response generation using an encoder-decoder model (illustration adapted from Li and Liang [2021]). Red colouring indicates that weights need to be tuned and stored. (Top) Fine-tuning one pre-trained model per persona. (Bottom) Tuning some additional weights (i.e. the encoder and decoder prefixes) for each persona which are prefixed to the pre-trained encoder and decoder. These remain unchanged.

difference among the presented studies and our work lies in how the persona representation is generated. We are the first to test prefixes as proposed by Li and Liang [2021] for personalised text generation.

4.2 Experimental Setup

In order to answer our RQs, we trained several models. In this section, we explain which models we trained for each RQ (section 4.2.1). We also present our customised training and inference setups (section 4.2.2) and the evaluation metrics (section 4.2.3) that we used to evaluate the results in chapter 5.

4.2.1 Experiments

This section lists the models that we train and links them to our RQs. Figure 10 offers a schematic view of all models.

The fine-tuned (FT) model gives us an intuition as to how well the traditional approach of adapting a large pre-trained model to a downstream task works.

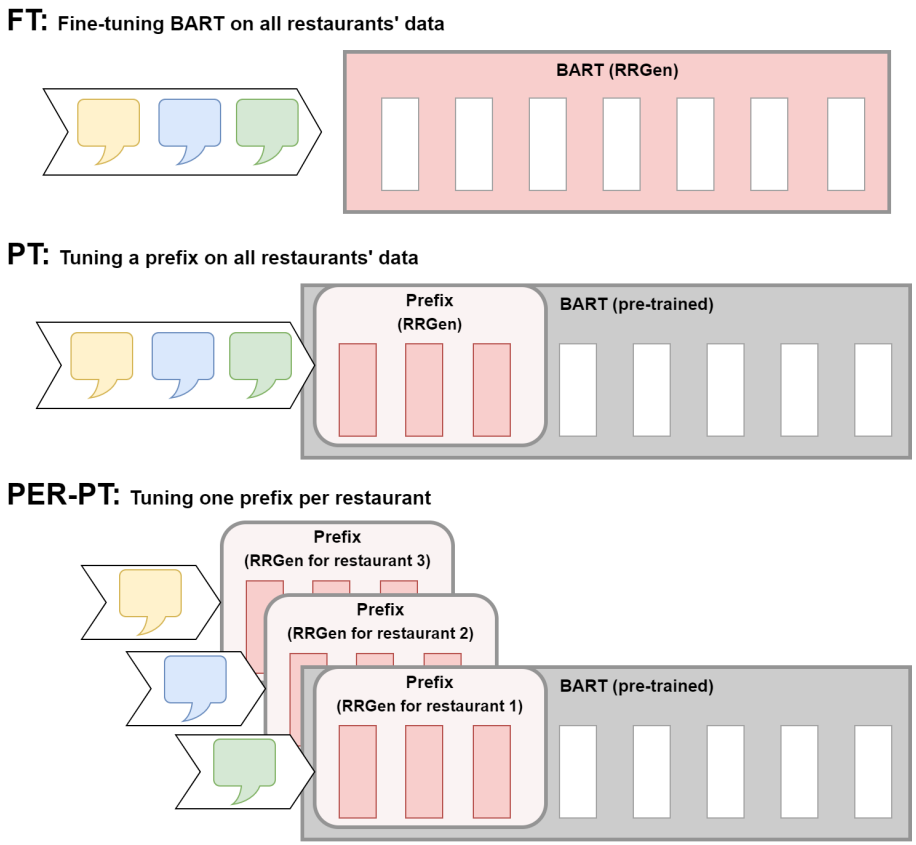


Figure 10: Tuned models and prefixes. Red colouring indicates that weights need to be tuned and stored. The fine-tuned (FT) model is the baseline model, i.e. the pre-trained model fully fine-tuned on the complete data set. The prefix-tuned (PT) model is the pre-trained model prepended with a prefix that is tuned on the complete restaurant data set. For the personalised prefix-tuned (PER-PT) model, we create a separate prefix for each restaurant in the data set and prefix it to the pre-trained model.

For **RQ1**, we investigate how well a prefix is suitable to adapt a pre-trained model to the downstream task of review response generation. To this end, we train the prefix-tuned (PT) model. We compare the outputs of this model to the outputs of the FT model to gauge to what extent the memory-friendly approach of prefix-tuning approximates the performance of fine-tuning in review response generation.

RQ2 inspects if prefixes are a suitable tool for reproducing a restaurant's style in the context of review response generation. To answer this question, we train the personalised prefix-tuned (PER-PT) model. We compare the generated responses of this model with those of the FT model to estimate if the task of review response generation is performed comparably well as by the baseline. Furthermore, we gauge whether the outputs are more characteristic of the restaurants in question.

A naive alternative to tuning personalised prefixes for meeting the requirement of personalisation would be to fine-tune a pre-trained model for each individual restaurant. However, given that only little training data is available for each restaurant, it is likely that the outcome would be unsatisfactory. Also, this approach results in as many large models as there are restaurants. Storing them would be highly memory-intensive. For these reasons, we do not test personalisation via full fine-tuning.

4.2.2 Model Training and Inference

Training

Our training is based on the implementation of Li and Liang [2021].[1] As our base model, we use the pre-trained model $\text{BART}_{\text{LARGE}}$. Li and Liang [2021] employed the same model for testing prefix-tuning on the summarisation task. BART is an effective and popular model to fine-tune towards different text generation tasks [Lewis et al., 2020]. BART is a Transformer-based model with a bi-directional encoder and an autoregressive decoder pre-trained on 160 GB of text data from different genres, such as books, news, and web-based text. Its pre-training tasks consisted of different forms of denoising corrupted input, such as reordering shuffled sentences and filling in masked tokens [Lewis et al., 2020].

We use the same hyperparameters for tuning the general review response prefix and the personalised prefixes, except for the different prefix lengths. The hyperparameters are listed in table 3. We set all of the hyperparameters equal to the default values as provided by the Hugging Face[2] library [Wolf et al., 2020] if not specified differently. We choose all other hyperparameters based on preliminary experiments and only tune explicitly for prefix length. How we do this is explained in section 4.3.

Different to Li and Liang [2021] and Zhao et al. [2019], and following Gao et al. [2019a] and Zhang et al. [2018], we use validation loss instead of a generation metric to find the best-performing models. The reasoning behind this is that the validation curves of the automatic reference-based metrics such as BLEU, ROUGE or chrF develop in a less stable manner than validation loss. Also, the development of the metrics' curves is often incongruent which makes their expressivity questionable.

The effective batch size for prefix-tuning results from multiplying the batch size we use, i.e. 7, with the number of gradient accumulation steps, i.e. 9. The effec-

[1] The source code can be found on X.L. Li's GitHub repository: `https://github.com/XiangLi1999/PrefixTuning` (last accessed 26 October 2022).

[2] `https://huggingface.co/` (last accessed 29 September 2022).

tive batch size used for fine-tuning is 126 (we use 18 gradient accumulation steps). Accumulating gradients over some steps means that we do not update the model parameters after every batch. Instead, we do as many backward passes with the same weights as the number of gradient accumulation steps. Only then do we update the weights with the gradients that we accumulated. This has the benefit that we can use effective batches that are larger than what fits on GPU memory.

Maximum source and target lengths are measured with the tokens from the BPE vocabulary [Sennrich et al., 2016] which is used by the BART tokeniser. With these maximum lengths, we found that 94.1% of the reviews and 94.0% of the responses can be processed by the model without being truncated.

Hyperparameter	(PER-)PT	FT
prefix length	200 (PT) 10 (PER-PT)	-
max. no. of epochs	30	30
early stopping patience	5	5
effective batch size	63	126
prefix parametrisation `mid_dim`	800	-
max. source length	256	256
max. target length	128	128

Table 3: Hyperparameter settings used during prefix-tuning and fine-tuning

For all tuning scenarios, we use the AdamW optimiser [Loshchilov and Hutter, 2017] and a linear learning rate scheduler.

The hyperparameter "prefix parametrisation `mid_dim`" refers to the size of the reparametrisation[3] network's intermediate layer. We choose the same value as Li and Liang [2021] chose for their summarisation task.

Model Inference

For decoding we generally use beam search with a beam size of 4.

Holtzman et al. [2019] found that sequence search algorithms that aim to maximise the output's probability, such as beam search does, tend to produce *de*generate text, i.e. boring, repetitive or incoherent text. Therefore, we also test top-p sampling

[3]Li and Liang [2021] noted that updating the prefix parameters directly causes the optimisation to be unstable. For more stability, they reparametrise the prefix parameter matrix with a smaller matrix passed through a feedforward neural network. During training, the smaller matrix and the network's weights are tuned. For inference, it suffices to store the final prefix.

(also known as nucleus sampling) [Holtzman et al., 2019] to generate more diverse responses. An advantage of top-p sampling is that we do not have to retrain the models in order to test it, as it only affects the decoding process.

Top-p sampling only samples from those words that constitute the top-p part of the probability mass at each decoding step [Holtzman et al., 2019]. The value p is set by the user. This kind of sampling ensures that samples are not drawn from the tail that contains only the least probable candidates. As the probability mass is fixed, the number of words to sample from changes dynamically at each decoding step. This differentiates top-p sampling from top-k sampling [Fan et al., 2018] where a sample from the top-k most probable words is drawn at each decoding step.

We set p to 0.9 and use three different seeds to generate responses. We average the metrics of the three inference runs to balance out potential variation between them. The results are in chapter 5.

4.2.3 Evaluation Metrics

The automatic evaluation of response generation is far from trivial [Kew and Volk, 2022; Zhang et al., 2019], let alone the evaluation of personalised response generation (cf. section 2.3 on an introduction of personalisation metrics). The open-endedness of response generation, i.e. the uncountable possibilities of how one can answer a review, makes reference-based metrics such as BLEU a less than ideal metric for response generation. Kew and Volk [2022] explicitly mention the need to include a variety of automatic evaluation metrics to measure different characteristics of the generated responses "along multiple axes". Due to the difficulty of assessing generated responses' quality, human evaluation is also widely used in studies of response generation [Gao et al., 2019b; Katsiuba et al., 2022; Kew and Volk, 2022; Ma et al., 2021; Smith et al., 2020; Su et al., 2019; Wang et al., 2017; Wu et al., 2021; Yang et al., 2018; Zhang et al., 2018, 2019; Zhao et al., 2019; Zheng et al., 2019]. Manual evaluation that can stand up to academic scrutiny is out of scope for this thesis. Especially for rating the degree of personalisation, there are no established automatic evaluation metrics but many proposals of less widely used metrics. Their usability for the task in question has not been scrutinised. Section 6.4 names a few problems of some metrics that we encountered during our work.

In the following, we list which automatic metrics we use to evaluate response generation and the degree of personalisation. If no further information is given, the usage of the metrics is the same as described in section 2.3.

- **chrF-tgt/chrF-src** From Kew and Volk [2022], we adopt the idea of calculating chrF against the input reviews to approximate the specificity of the generated responses. The rationale behind this is that more specific responses pick up topics from the input and thus repeat n-grams from the reviews. We denote these scores with the suffix "-src". "-tgt" means that the ground truth responses are used as the reference.
- **BLEU-tgt**
- **IntraDIST-1** Distinct-N [Li et al., 2016a] within a response, as proposed by Kew and Volk [2022]
- percentage of **unique responses** and number of **unique words**
- **Self-BLEU**
- **average text length**
- **perplexity** We interpret the scores in another way as previous studies on personalised response generation [Li et al., 2016b; Smith et al., 2020; Wang et al., 2021; Wu et al., 2021; Yang et al., 2018; Zhang et al., 2018; Zheng et al., 2019]: While these studies strived to minimise the perplexity of the models' generated outputs, we think that achieving a perplexity close to the ground truth's is a more suitable goal. Those who minimise perplexity argue that lower perplexity is an indicator of more grammatical and fluent text [Jin et al., 2022; Wang et al., 2021; Wu et al., 2021; Zheng et al., 2019]. We do not argue against this reasoning, but Holtzman et al. [2019] found that texts with very low perplexity tend to have too low diversity which may indicate that the texts are bland. For this reason, they consider texts that are closest to ground truth's perplexity as best w.r.t. diversity while still being grammatical.
- **restaurant classifier accuracy** Our model is a `fastText`[4] classifier [Joulin et al., 2017]. `fastText` learns word embeddings during training. Each input text is represented by its averaged word embeddings and its bag of n-grams as an additional feature. This hidden representation enters a linear classifier. For each restaurant, we trained a one-versus-all classifier for 50 epochs with a learning rate of 0.3 on the combined training and validation sets of our data set. In order to include some local word order information, we used bags of trigrams as an additional feature. The model achieved a micro-averaged accuracy of 91% on the test set, the class accuracies ranged from 55.3% to 99.4%. Since some of the responses contained restaurant names which would

[4] `https://fasttext.cc/` (last accessed 13 September 2022).

make classification overly simple for the model, we masked all restaurant names from both the training and the test sets. We do the same in the generated responses prior to inputting them to the classifier.

- **MaxBLEU** and **MaxBLEU accuracy**

Generally, we want to achieve metric values that are as close as possible to the metrics of the ground truth responses. BLEU-tgt and chrF-tgt are exempt from this rule: We want to maximise these two metrics.

Since we aim to generate responses with a similar amount of diversity and "surprisingness" as the human-written responses, our goal is to achieve similar chrF-src, IntraDIST-1, numbers of unique responses and words, Self-BLEU, text length and perplexity as in the ground truth responses.

Also regarding the personalisation metrics, we strive to obtain similar scores for the generated responses as for the reference responses. The reason why higher personalisation scores are not necessarily better is that not all responses contain distinctive characteristics that make an attribution to an author possible. Therefore, 100% accuracies are not achievable if we want to generate responses that are similar to human-written responses.

4.3 Prefix Length: Preliminary Experiments

Li and Liang [2021] analysed how changing the prefix-tuning hyperparameters affected the results. They found that increasing the prefix length up to a certain point improved the outputs, while going beyond this point resulted in a slowly deteriorating performance. On a more general level, they also discovered that the optimal prefix length is task-specific. These findings suggest that tuning for the optimal prefix length for both general review response generation and personalised review response generation is a decisive step in deploying prefix-tuning effectively. The following sections describe the effect of prefix length on different aspects of review response generation using both the PT and the PER-PT models. Except for the prefix length, we used the same hyperparameters as indicated in section 4.2.2.

4.3.1 Tuning Prefix Length for Downstream Task Adaptation

The goal of the first experiment is to find the optimal prefix length for general review response generation with the PT model. To do so, we tuned seven prefixes with the

lengths in {1, 10, 25, 50, 100, 200, 300}.[5] Since the model's aim is to solve the downstream task of review response generation in general, we used the entire data set for these hyperparameter tuning runs. Figure 11 shows four different metrics of the tuned models: We see that the two shortest prefix lengths 1 and 10 result in the worst metrics across the board. With a prefix length of 25, all metrics, especially the loss, improve visibly. If we consider the reference-based metrics ROUGE-1 and BLEU, prefix lengths of 50 or 200 result in the best metrics. Meanwhile, loss is lowest for 100 or 200. At the prefix length of 300, all metrics except for text length start to deteriorate slightly.

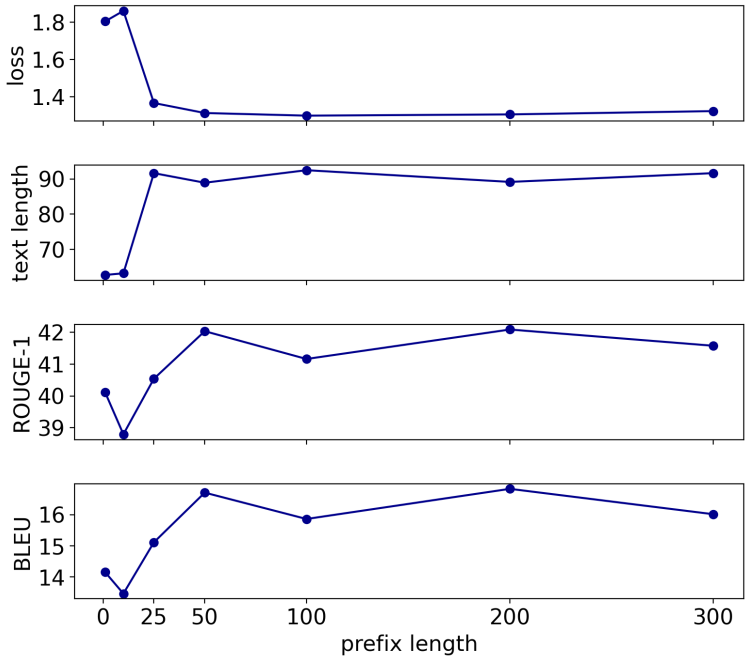


Figure 11: Validation metrics of PT model with different prefix lengths. Note: Text length is measured in byte-pair encoded subword units generated by BART's tokeniser.

To select the optimal prefix length, we also consider the diversity of the generated responses using different prefix lengths. Figure 12 shows that the Self-BLEU values stay relatively consistent across the prefix lengths ranging from 25 to 300. This indicates that from a certain prefix size onwards, prefix length has only a marginal effect on the generated responses' diversity.

[5]Li and Liang [2021] selected their prefix length for the summarisation model from a range of values between 1 and 300 as well.

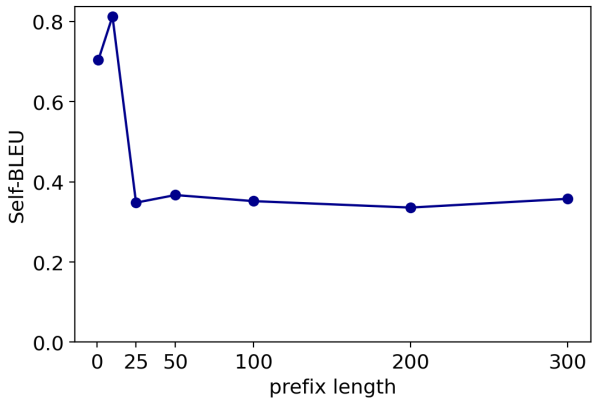


Figure 12: Diversity of validation set responses generated by PT model with different prefix lengths (averaged across all restaurants)

We have seen, analogously to Li and Liang [2021], that the performance improves with increasing prefix size up to a certain point and slowly starts to deteriorate with larger prefixes. For our subsequent experiments, we opt for a prefix length of 200 for the PT model, as ROUGE-1, BLEU, and diversity are highest for our data and task when using this prefix length. We report the results of the experiments with the PT model in section 5.1.

4.3.2 Tuning Prefix Length for Personalisation

To find the optimal prefix length for personalised review response generation, we tuned some prefixes on a subset of the data set. Specifically, we selected five restaurants that are different regarding various aspects as can be seen on table 4: They serve food from different cuisines in varying price categories and four countries around the world, and they have differing numbers of review-response pairs. In fact, "Hard Rock Cafe" is the restaurant with the most review-response pairs in the data set, whereas "Pizza Pilgrims" is the one with the least. In case that the optimal prefix length varies across restaurants, this selection will disclose such a tendency.

We tested prefix lengths in {1, 10, 25, 50, 100, 200} by tuning a prefix for every combination of restaurant and prefix length. The training setup was the same as indicated in section 4.2.2. Figure 13 summarises the results of these hyperparameter tuning runs by showing the values of four validation metrics. Each line on a plot marks the performance for one of the five restaurants. Overall, the impact of the prefix length on the generation results is moderate, as most lines are relatively horizontal across the different prefix lengths. In the following, we discuss some trends and special behaviours.

Restaurant	Cuisine	Place	Price category	Avg. resp. length (tokens)	RR Pairs
Hard Rock Cafe	American	Orlando, USA	“mid-range”	20	5,854
Salt House	Contemporary	Cairns, Australia	“mid-range”	48	2,472
Cinnamon Club	Indian	London, England	“fine dining”	68	2,254
Red Torch Ginger	Asian	Dublin, Ireland	“mid-range”	73	1,369
Pizza Pilgrims	Italian	London, England	“cheap eats”	49	1,193

Table 4: Selected restaurants for tuning length of personalised prefixes

Plot 13.a) depicts the validation loss for each prefix model. The lines of “Pizza Pilgrims”, “Salt House” and “Hard Rock Cafe” show that the loss is lowest around a prefix length of 10 or 25. The lines of “Cinnamon Club” and “Red Torch Ginger” also have their lowest validation loss at either 10 or 25. However, unlike the other three restaurants, their losses do not grow monotonically from 25 onward. This indicates that the minimal losses we discovered for these restaurants may not be the globally minimal loss values. Nevertheless, when averaging over all restaurants, validation loss is lowest at a prefix length of 10. This is why we tune all following prefixes for personalisation with this prefix length.

The curves’ shape of the reference-based metrics BLEU and ROUGE-1 in plots 13.c) and 13.d) are similar, but the ordering of the restaurants is different: For example, the prefixes of “Hard Rock Cafe” generate better results than the prefixes of “Pizza Pilgrims” w.r.t. BLEU, while the outcomes are reversed for ROUGE-1. We assume this difference stems from BLEU being a precision-based metric, whereas ROUGE is based on the n-gram recall. As the generated responses of “Hard Rock Cafe” are shorter than those of “Pizza Pilgrims” (cf. figure 13.b), it is harder for these shorter texts to achieve a high recall. When looking at peaks in the BLEU and ROUGE curves, only “Cinnamon Club” and “Hard Rock Cafe” have a maximum around 10 or 25, while the curves of “Salt House” and “Red Torch Ginger” fluctuate more. The curves of “Pizza Pilgrims” are relatively flat without having discernible maximums. Looking at the average curves, we see that the highest BLEU and ROUGE-1 scores are achieved with a prefix length of 10, although for ROUGE-1, the differences between the reported values for prefix lengths between 10 and 100 are marginal.

The differences in text generation length in plot 13.b) are the most pronounced between the prefix lengths, although the behaviour across the restaurants is different. “Cinnamon Club” shows a drastic increase in generated text length from a prefix

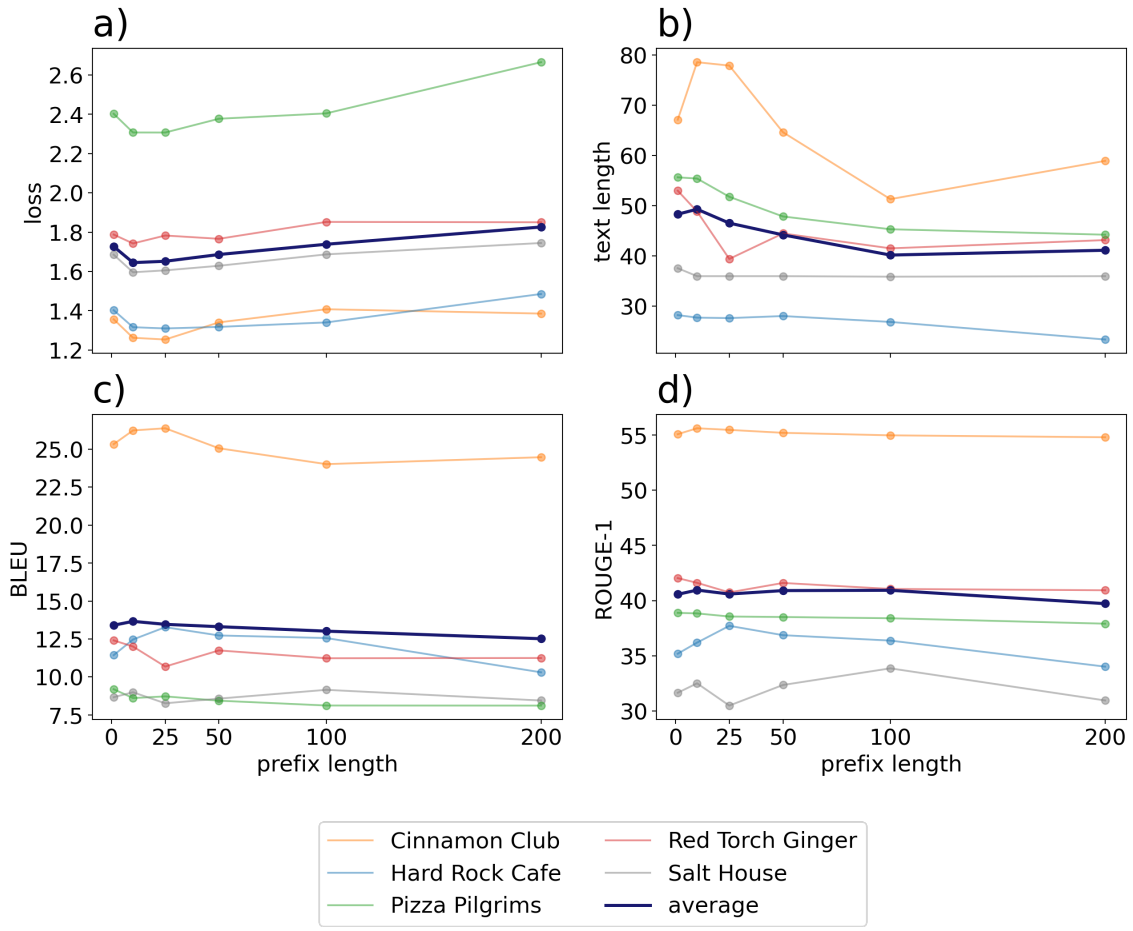


Figure 13: Validation metrics of PER-PT model with different prefix lengths. Note: Text length is measured in byte-pair encoded subword units generated by BART's tokeniser.

length of 1 to a prefix length of 10, and a drastic decrease of text length from a prefix length of 25 to a prefix length of 50. This indicates that the longest texts are generated with a prefix length around 10 and 25 in the case of "Cinnamon Club". The differences are less distinct for the other restaurants: The generated texts of "Pizza Pilgrims" and "Red Torch Ginger" are highest for short prefix lengths of 1 or 10. For the restaurants "Salt House" and "Hard Rock Cafe", the generated texts vary only slightly in length across the different prefix lengths. We see a reason for this in the short length of the training set responses for these restaurants (cf. table 4) which makes it unlikely that the trained models start generating longer texts for these restaurants without an incentive from the training set.

We also compare the diversity of the generated responses across prefix lengths. Figure 14 shows the Self-BLEU scores for the selected restaurants and prefix lengths.

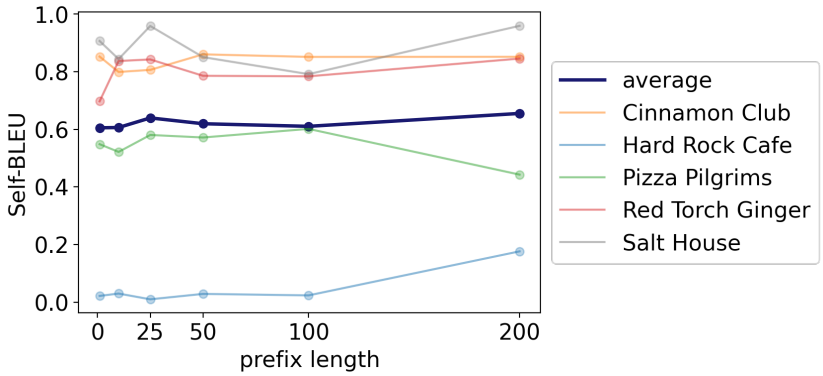


Figure 14: Diversity of validation set responses generated by PER-PT model with different prefix lengths (averaged across all restaurants)

There is no clear trend in diversity with increasing prefix lengths, as each restaurant's Self-BLEU (except for "Hard Rock Cafe") as well as the average Self-BLEU fluctuate across prefix lengths.

Based on the results of these preliminary experiments, we set the prefix length to 10 for our experiments with the PER-PT model. We have shown that, overall, prefix length has a limited effect on the model performance and the diversity of the generated responses in the case of personalised response generation. The next chapter presents the results of the trained review response generation models on our TripAdvisor data set using the prefix lengths that we selected in this section.

5 Results

In this chapter, we present the results of the trained review response generation models described in section 4.2.1. Section 5.1 focuses on RQ1, as we first compare both models trained to solve the general task of review response generation. The goal of this is to assess how well prefix-tuning is suited to learn this downstream task compared to fine-tuning. In section 5.2, we evaluate the personalisation capabilities of prefix-tuning to answer RQ2. Section 5.3 contrasts the generated responses with the human-written ground truth responses, and section 5.4 compares our results with those of previous work.

In tables 5 and 6, the numbers closest to the ground truth are marked in **bold** except for BLEU-tgt and chrF-tgt where the highest (and best) values are **bold**. Endings "-tgt" and "-src" mean that the target responses and the input reviews were the reference respectively. We calculated the metrics on each restaurant's test set individually and aggregated them over all restaurants to produce the macro averages reported here. Whenever applicable, we multiplied the values with 100 to improve readability. Since we used three different seeds to generate the responses with top-p sampling, we report the average metrics with their standard deviations in parentheses.

5.1 Learning a Downstream Task: Prefix-Tuning vs. Fine-Tuning

In this section, we compare the FT and PT models which are illustrated in figure 10 (page 34). Fine-tuning produces our baseline model for review response generation, whereas prefix-tuning is the experimental approach to be tested. Table 5 shows the evaluation metrics (described in section 4.2.3) of the FT and PT models.

We first compare the results for the FT and the PT models using beam search decoding: The higher BLEU-tgt and chrF-tgt scores indicate that the FT model's responses are closer to the ground truth sentences than the PT model's responses.

	Metric	Ground truth	FT (BS)	FT (top-p)	PT (BS)	PT (top-p)
General	BLEU-tgt ↑	-	**18.5**	14.9 (± 0.37)	16.5	14.1 (± 0.31)
	chrF-tgt ↑	-	**37.4**	34.7 (± 0.36)	35.5	33.4 (± 0.17)
	perplexity	94.3	63.9	42.2 (± 1.86)	64.1	**64.6** (± 10.76)
Diversity/ specificity	chrF-src	17.4	14.8	**16.0** (± 0.27)	14.1	14.8 (± 0.18)
	intraDIST-1	79.8	80.4	**79.7** (± 0.31)	81.6	80.4 (± 0.14)
	Self-BLEU	9.3	37.0	**13.7** (± 2.14)	37.2	17.2 (± 0.64)
	unique responses	92.0	51.0	**84.7** (± 0.84)	32.8	74.2 (± 1.44)
	unique words	987.2	414.7	**758.2** (± 7.71)	330.2	632.7 (± 6.29)
	avg. text length	58.3	50.1	**52.1** (± 0.83)	49.7	49.3 (± 0.69)

Table 5: Evaluation metrics of FT and PT models using beam search (BS) and top-p sampling (top-p) decoding, and ground truth. For better visibility, columns with results for top-p sampling are shaded light grey.

The perplexity scores of both models are similar, suggesting that both models produce responses that are considered equally likely by a language model.

The chrF-src scores indicate that the FT model is more likely to pick up n-grams from the input review than the PT model, thus responding slightly more directly to the topics in the review. The intraDIST-1 scores are similar for both models and the ground truth responses. This indicates that the intra-textual diversity is comparable for them. In other words, both models do not get caught in undesired repetitive loops. The Self-BLEU scores of both models are similar too, so both models repeat n-grams across the generated responses to a similar degree. The FT model's generated texts, however, are characterised by a larger inter-textual diversity, as they contain more unique responses and words than the PT model's produced texts. The average text length is similar for both models. To sum up, the FT model produces more diverse and slightly more specific responses than the PT model.

The FT and PT models generate texts that are less diverse than the human-written responses when using the standard decoding strategy, i.e. beam search. In order to elicit more diverse responses from the models, we leverage top-p sampling[1] [Holtzman et al., 2019]. In summary, the reference-based metrics deteriorate when using the FT and PT models with top-p sampling rather than beam search. chrF-src increases which means that the generated responses are slightly more specific. The texts are slightly longer in the case of the FT model. The diversity increases notice-

[1] A description of top-p sampling can be found in section 4.2.2.

ably for the FT as well as the PT model such that it is closer, though not equal to the diversity in the ground truth responses. While greater inter-textual diversity is a desirable property for the review responses, a sampling-based decoding approach increases the risk of unwanted behaviour such as the model hallucinating incorrect facts. We discuss the topic of hallucination more closely in section 6.3.

All in all, the PT model approximates the performance of the FT model in all respects expect for the rates of unique responses and words. Thus, the FT model produces more diverse responses than the PT model. We found that the diversity of both models' responses can be increased by using top-p sampling in place of beam search with small losses for the reference-based metrics. Given that prefix-tuning is considerably more lightweight than fine-tuning w.r.t. storage space, prefix-tuning is indeed a viable alternative to fine-tuning for general review response generation when memory efficiency is crucial.

5.2 Learning Styles: Personalised Prefix-Tuning vs. Fine-Tuning

This section addresses personalised review response generation using prefix-tuning. In order to measure whether prefix-tuning is a suitable tool for reproducing different restaurants' review response characteristics, we compare the FT model with the PER-PT model. Both are shown in the overview in figure 10 (page 34). Table 6 compares the performance of the two models (metrics are described in section 4.2.3).

We first inspect the results for the responses generated with beam search decoding: The BLEU and chrF values calculated on the human-written references indicate that the PER-PT model outperforms the FT model w.r.t. n-gram overlaps with the ground truth. This is not surprising, as the PER-PT model's training data was less diverse than the FT model's training set: The first consisted of single restaurants' data, whereas the second encompassed data from all restaurants. A large uniformity in the training data increases the probability that the generated responses have a higher n-gram overlap with the ground truth responses. The perplexity scores of the two models are similar to the ones already calculated for the PT model in the last section. Thus, the generated texts of all of our trained models are almost equally expectable from the perspective of a language model.

Regarding specificity, the models also perform comparably to the PT model from the last section with similar chrF-src scores. IntraDIST-1 is similar to the results in the last section too. Therefore, the models in this section do not get stuck in

	Metric	Ground * truth	FT * (BS)	FT * (top-p)	PER-PT (BS)	PER-PT (top-p)
General	BLEU-tgt ↑	-	18.5	14.9 (± 0.37)	**21.1**	17.6 (± 0.38)
	chrF-tgt ↑	-	37.4	34.7 (± 0.36)	**39.0**	36.1 (± 0.29)
	perplexity	94.3	63.9	42.2 (± 1.86)	63.5	**71.3** (± 3.12)
Diversity/ specificity	chrF-src	17.4	14.8	**16.0** (± 0.27)	14.0	14.6 (± 0.02)
	intraDIST-1	79.8	80.4	**79.7** (± 0.31)	80.7	80.7 (± 0.16)
	Self-BLEU	9.3	37.0	**13.7** (± 2.14)	64.5	23.0 (± 1.35)
	unique resp.	92.0	51.0	**84.7** (± 0.84)	35.2	75.4 (± 10.6)
	unique words	987.2	414.7	**758.2** (± 7.71)	158.5	485.2 (± 26.2)
	avg. text length	58.3	50.1	**52.1** (± 0.83)	47.1	47.6 (± 0.07)
Personali- sation	MaxBLEU	53.7	62.7	44.6 (± 0.48)	79.1	**57.8** (± 0.74)
	MaxBLEU acc.	84.1	60.2	48.5 (± 0.52)	92.7	**78.9** (± 0.45)
	rest. class. acc.	91.4	64.0	46.1 (± 0.19)	98.4	**94.5** (± 0.57)

Table 6: Evaluation metrics of FT and PER-PT models using beam search (BS) and top-p sampling (top-p) decoding, and ground truth. For better visibility, columns with results for top-p sampling are shaded light grey.

repetitive loops as well.

The other scores from the category of inter-textual diversity, however, indicate large differences between the models' responses. The Self-BLEU scores show that there are many more n-gram repetitions between the generated responses of the PER-PT model than of the FT model. The fraction of unique responses and the number of unique words speak a similar language, showing that the responses of the PER-PT model are more homogeneous than those of the FT model. Therefore, the PER-PT model produces more repetitive responses than the FT model.

Although the responses of the PER-PT model are highly repetitive (and probably generic), they are characteristic of the restaurants for which they are produced. This is indicated by the personalisation metrics which score well.

As in the previous section, both the FT and the PER-PT models produce less diverse responses than ground truth responses when using beam search for decoding. Therefore, we again use top-p sampling [Holtzman et al., 2019] for decoding to encourage the models to produce more creative responses. The results are in table 6. BLEU and chrF against ground truth decreased compared to the beam search de-

*These columns are identical to the ones in table 5.

coded responses. However, for all other metrics, top-p sampling resulted in outputs that are, on average, closer to the ground truth for the PER-PT model: The generated responses are more inter-textually diverse, have a greater lexical range and are more "surprising" (as perplexity is higher). For the FT model, top-p sampling results in more diverse but less personalised responses than if using beam search.

In summary, the PER-PT model successfully learned to produce responses that are more characteristic of the restaurants than the responses generated by the FT model, as its personalisation scores exceed the FT model's. Its responses are also closer to the reference responses than those of the FT model. Only when considering inter-textual diversity does the FT model's performance surpass that of the PT model, but this disparity is resolvable through top-p sampling.

5.3 Comparing Generated With Reference Responses

We have seen in the previous two sections that the PT and PER-PT models lag behind the FT model when using beam search in terms of inter-response diversity and lexical range and to a small extent also w.r.t. specificity. We also showed that the lack of diversity and specificity is somewhat improved using top-p sampling. This comes at the cost of small performance drops w.r.t. the other metrics for the PT and PER-PT models. The lowered personalisation scores for the PER-PT model, however, are unproblematic, as they moved closer to the ground truth's scores.

In this section, we will compare the responses generated by our models with the ground truth responses in terms of diversity, specificity, and personalisation.

5.3.1 Diversity and Specificity

Considering tables 5 and 6, we see that all of our trained models produce responses with less inter-textual and lexical diversity than the human-written responses: The rates of unique responses and counts of unique words are smaller and the Self-BLEU scores are larger for the model outputs than for the ground truth responses. The tables also show that these differences persist when using top-p sampling, although to a smaller extent. Regarding specificity, the chrF-src values are lower and the average text lengths of the generated responses are shorter than those of the human-written responses. Thus, the generated responses are less specific to the input reviews than the ground truth responses. This applies to the responses decoded with beam search as well as with top-p sampling. Finally, the perplexity of the model-produced re-

sponses is lower across the board, meaning that these responses are less "surprising" than the ground truth responses.

These observations suggest that the models overproduce low-diversity, low-specificity and short responses compared to the references. These properties are symptomatic of "one-size-fits-all" responses, i.e. responses that can be re-used for a large array of different reviews. The tendency of producing such bland responses has been observed for different neural text generation models [Holtzman et al., 2019; Kew and Volk, 2022]. We show examples of such universal responses in the discussion's section 6.1.1.

5.3.2 Personalisation

In this section, we compare the personalisation metrics for the generated responses of the FT and PER-PT models and the human-written responses.

Table 6 shows that both the FT and PER-PT models using beam search have higher MaxBLEU scores than the ground truth. This means that the models' generated responses are more similar to the training set responses than the ground truth test set responses. In other words, the models learned to a disproportionately high extent to copy n-grams from the training set compared to the ground truth responses.

Regarding MaxBLEU accuracy and restaurant classifier accuracy, the pattern is different: Table 6 shows that the PER-PT model has higher accuracy scores than the ground truth responses, while the FT model has the lowest scores for these metrics using beam search. The low MaxBLEU accuracy of the FT model indicates that it generates responses that are less similar to the respective restaurant's training set compared to the PER-PT model's responses and the ground truth responses.

When top-p sampling is used as the decoding strategy, the personalisation scores of the PER-PT model approximate the scores of the ground truth responses as can be seen on table 6. This indicates that producing more diverse responses minimises the degree of personalisation. However, enough of the characteristic properties are retained in the sampling-based decoding approach to achieve similar personalisation scores as the ground truth responses.

All in all, the PER-PT model produces review responses that are at least as characteristic of the respective restaurants as the ground truth responses. The FT model's responses are less personalised towards individual restaurants.

5.4 Comparing Our With Previous Studies' Results

We contrast our work with previous research that is similar to ours, although the results are not directly comparable because of varying experimental setups. Considering the BLEU and chrF values, our trained models outperform previous results. Kew et al. [2020], who used a basic attentional seq2seq model on a four times larger data set of review-response pairs from restaurants and hotels, achieved 8.17 BLEU points. Kew and Volk [2022], who used around 2.25 times more review-response pairs from hotels with highly generic responses filtered out to fine-tune $\text{BART}_{\text{BASE}}$, achieved a maximum chrF score of 33.6. We assume that our chrF and BLEU values are higher than in the previous studies because we used $\text{BART}_{\text{LARGE}}$ which is a larger and thus more powerful model that is pre-trained on more data. Another difference between our and the previous studies lies in the pre-processing step of masking greetings and salutations with tags. Predicting tags is easier for the model than predicting greetings and salutations which may force the model to hallucinate names that have not been mentioned in the reviews, resulting in lower scores.

Kew and Volk [2022] achieved a chrF-src score of 20.7 using their proposed training data filtering methods. Therefore, the specificity of their generated responses is higher than the specificity of our models' responses (and also higher than the specificity of our ground truth responses). A difference between this study and the study of Kew and Volk [2022] is that they exclusively worked with data from hotels, while we used review-response pairs from restaurants. The data set descriptions differ as well, as the average text length of their data set is considerably longer than ours, and their Self-BLEU score is lower. Therefore, this comparison between the studies should be taken with a grain of salt.

6 Challenges in (Personalised) Review Response Generation

For our third research question, we aim to identify some of the challenges of review response generation. For this purpose, we conduct a qualitative analysis of a few examples from the TripAdvisor data set and the generated texts of our trained models.[1] While doing so, we compare the trained models in different aspects. Thus, the findings from this chapter also illuminate the strengths and weaknesses of the individual models. This is helpful for further differentiating our answers to RQs 1 and 2. In section 6.1, we study responses with varying inter-textual diversity and how this relates to specificity. We also analyse the effect of top-p sampling on response quality. In section 6.2, we investigate how our models react to negative reviews. In section 6.3, we quantify the occurrences of two types of hallucination in the generated responses. In section 6.4, we discuss some issues related to existing personalisation metrics.

6.1 Diversity and Specificity

In this section, we manually inspect some responses generated by the PER-PT model to discover the characteristics of its high-frequency responses and its strategies for producing more diverse and specific responses. While the first two subsections, 6.1.1 and 6.1.2, focus on responses generated using beam search, we dedicate subsection 6.1.3 to top-p-sampled responses.

[1] The examples from the data set and the generated responses are reproduced in this chapter as they are. We did not edit any of texts to remove orthographic or grammatical errors. The only alterations we made include italicisation of text parts to highlight the most relevant passages for the discussion and removal of irrelevant text passages (marked by "[...]").

6.1.1 Low-Diversity Responses

The PER-PT model suffers from low inter-textual diversity when run with beam search decoding (cf. section 5.2). In the most extreme cases, the model produces only two different outputs for all reviews of a restaurant's test set, as is the case for the restaurant "Donatello". The generated responses are given in examples 3 and 4.

(3) Thank you for your review; we are glad you enjoyed your experience and we hope to see you again soon!

(4) Thank you for your review; we are sorry you didn't find your experience with us enjoyable but we will take your comments on board.

Both responses are highly universal, as they do not address any compliments or issues raised in the reviews directly. The tuned prefix learned two different ways of reacting to reviews: one suitable for positive reviews and one for negative ratings. Unsurprisingly, these two responses also occur repeatedly in the training examples for this restaurant: The first constitutes 7.5% of the training set, the second 1.9%. Thus, both are relatively frequent. Provided that the model uses the positive and negative responses appropriately, the model learned to detect sentiments. While this is useful, a simpler model requiring less computing power for training and inference could have accomplished this task as well.

For the Indian restaurant "Urban Tandoor", the PER-PT model also generates only two different responses which are part of the training set as well. However, both responses refer to a positive experience. In this case, the PER-PT model did not learn to differentiate between positive and negative reviews.

A short discussion on how overly general review responses may affect the eWOM of a business can be found in section 6.2.1.

6.1.2 High-Diversity Responses

In sections 5.1 and 5.2, we saw that the PT and FT models produce more diverse responses than the PER-PT model when using beam search. We hypothesise that much of the diversity of these models stems from the larger diversity in the training set, as these models were trained with all restaurants' data. While larger training sets offer more knowledge to the models, learning information that does not fit all restaurants (e.g. about restaurant-specific dishes) leads to the models hallucinating and making false claims as we will discuss in section 6.3. In this section, we will

only study sources of diversity that do not arise from the models hallucinating. The illustrative examples in this section were produced by the PER-PT model.

One way of achieving more diversity and specificity is by copying elements from the reviews. Frequently copied mentions include meals, dishes, waiters' names, or special menus, such as in example 5 where they are marked in italics.

(5) a. **Review:** [...] From the many choices on the *A La Carte Menu* we chose the Rack of Lamb & Duck main dishes [...]. The Service by *John & Claire* was very attentive [...].

b. **Generated response:** Thank you so much for taking the time to share your recent dining experience with us, we are so happy to hear that you enjoyed our *a la carte menu* and the service from *John & Claire*! [...]

The PER-PT model learned to pick up entire compliments by the customers and to repeat them as in example 6.

(6) a. **Review:** [...] *The Brahma on draught is great and compliments the meat very well* in my opinion. [...]

b. **Generated response:** [...] I am so pleased to hear you enjoyed your dining experience with us. *I agree, the Brahma on draught is great and compliments the meat very well.* [...]

Example 7 shows that the models sometimes copy idiomatic expressions such as "to go the extra mile" to achieve specificity towards the input review, although the results sound unnatural. We assume that humans do generally not copy expressions from reviews, but rather paraphrase them as in the following example that occurs in the training examples in a similar way: "We are so happy that Nobu's outstanding service could make your birthday celebration such a memorable one [...]". Example 7 also illustrates that congratulating on someone's birthday is another way of making responses more specific to the review.

(7) a. **Review:** What a way to celebrate *my birthday*! [...] outstanding service by *Nobu who went the extra mile* making our dining experience an unforgettable one! [...]

b. **Generated response:** *Happy belated birthday to you!* We are so happy to hear that you had such an unforgettable experience with us and that *Nobu went the extra mile* for you. [...]

The customers' descriptions of the food is sometimes copied into the responses as well, as in example 8. However, this example shows that the model sometimes fails to retain the correct reference of the descriptions in the generated responses: Actually, the customer refers to the fish cakes and scallops as "bland and boring", but the generated response assigns this description to the barramundi.

(8) a. **Review:** The *barramundi main* was ok the rest ranged from *bland and boring* fish cakes and scallops to *horrid*, cheap tasting *desserts*. [...]

b. **Generated response:** [...] We are sorry to hear that you found the *barramundi main* to be *bland and boring* and the *desserts* to be *horrid*. [...]

To sum up, the different ways of generating more diverse responses mostly include copying content from the input reviews. This also results in responses that are more specific towards the input reviews.

6.1.3 Effect of Decoding With Top-*p* Sampling

We have seen in chapter 5 that top-p sampling helped the models to generate more specific and diverse responses. In this section, we illustrate these findings with some of the generated responses.

Example 9 shows how decoding with top-p sampling results in a more specific response than decoding with beam search. Three elements of the review, i.e. the waiter, the band, and the food, occur in the top-p-sampled response. Meanwhile, the response generated using beam search mentions no aspects specific to the review.

(9) a. **Review:** [...] Main waitor *KRIS* was outstanding. [...] The *band* SPEAKEASY we brilliant. [...] Forgot to mention the *food*. Steaks were the best I have ever had. Brilliant night [...]

b. **BS response:** Thank you so much for your kind review. We hope it won't be too long before your next visit as we would love to welcome you back soon.

c. **Top-*p* response:** Thank you for your kind review. I will pass your comments to our executive chef, *KRIS*, so we can look into them further. I am glad to hear that you enjoyed the *band* and *eating* with us. [...]

Another beneficial behaviour that we observed for top-p sampling was that it helped the model to detect negative reviews as in example 10. While the response produced

with beam search does not address the issues raised in the review, the sampled response apologises for the problems and paraphrases the described event. The superiority in identifying bad reviews when using top-p sampling compared to beam search is also addressed in section 6.2.2.

(10) a. **Review:** [...] The guy running the door was so so rude and smug, not a nice human being. He tried to charge us to get in when others were let in free, and then gloated when we walked away. [...]

b. **BS response:** Thank you so much for your kind review. We hope it won't be too long before your next visit as we would love to welcome you back soon.

c. **Top-p response:** Many thanks for your review. I am *sorry* to hear that you felt you were *disadvantaged* by this one particular experience and I hope we can bring you back soon.

While it is desirable that the model not merely copies words and sentence fragments from the review but rather paraphrases the aspects addressed in them, clumsy paraphrases result in unnatural or impolite responses as in example 11. In this example, the model identifies the issues named by the reviewer, but it addresses them in an unconventional way.

(11) a. **Review:** [...] Be aware of the *hidden charges* like the bread , we weren't expecting that. When asked if we could *add an additional team member* to the party of 24 with 24 hours notice, it took a lot of persuading staff. Disappointing[...]!

b. **Top-p response:** I am sorry to hear that you felt disappointed[...]. As you said on the review, our team *can't afford to overbook* and you should be assured that we *strive to keep costs under control* where possible. [...]

In some cases, the increased amount of creativity also leads to the model producing incomprehensible or contradictory responses as in examples 12a and 12b.

(12) a. I am sorry to hear you weren't pleased with the service, you would not be allowed to return for a second visit because we strive to be one of the very best in the city. We have taken the liberty to accommodate your feelings and will rectify any differences we can in the near future.

b. I'm *sorry* to hear that you had a *good* experience with us.

Based on these examples, we see that using top-p sampling helps to recognise negative reviews, and it leads to more interesting and specific responses with more paraphrasing. In some cases, however, encouraging more creative, less probable responses results in the model producing abstruse formulations.

6.2 Responding to Negative Reviews

Reviews on platforms such as TripAdvisor influence consumers' decisions [Zhang and Vásquez, 2014]. Negative reviews can be especially hurtful for a business, as they are considered more credible by online readers [Levy et al., 2013]. It has been found that ignoring negative reviews in online reputation management has a negative effect on a company [Chan and Guillet, 2011], while providing a response increases perceptions of a company's trustworthiness [Sparks et al., 2016]. Thus, responding appropriately to negative reviews is an important, although challenging aspect of managing a company's eWOM [Sparks et al., 2016]. A guide published by TripAdvisor recommends responding to all reviews within a day and prioritising negative reviews [Barsky and Frame, 2009]. While (semi-)automatic response generation helps to alleviate the pressure of responding to reviews in a timely fashion, it is still necessary to handle complaints appropriately and with tact.

Given that responding effectively to negative reviews is so important for a company's eWOM, we investigate how our models handle criticism. Generally, the prerequisites for generating adequate responses are suboptimal: On the one hand, review-response pairs with negative ratings are underrepresented in our data set. Figure 7 (page 29) shows that 1- and 2-star ratings only make up 7.8% of our data set. This complicates learning an appropriate behaviour towards complaints. On the other hand, reacting to criticism needs more finesse and world knowledge compared to compliments. Levy et al. [2013] recommend listing corrective actions for solvable issues and explanations for unsolvable problems. For both, world knowledge or even insider knowledge of a company are required. These factors make it especially difficult for an automatic system to react appropriately to negative reviews.

In section 6.2.1, we first identify some common but suboptimal strategies used by the models for replying to negative reviews. We also discuss how these strategies differ from practices found in human-written responses. In section 6.2.2, we approximate the recall of negative reviews of our trained models using a heuristic approach.

6.2.1 Strategies for Automatically Dealing With Criticism

In this section, we qualitatively analyse the generated responses to identify some of the most common strategies used by the models for reacting to negative reviews. We consider these strategies suboptimal since they diverge from the best practices listed in section 6.2.2. All examples shown in the following were generated by the PER-PT model using beam search.

We identified the following strategies as some of the most frequently used by the trained models to respond to negative reviews:

- responding in a highly unspecific manner,
- copying criticisms from the reviews without explanations or relativisations,
- hallucinating complaints,
- copying long text sequences from the review to generate "awkward" responses,
- and not reacting to criticism at all.

The first strategy results in **responses that lack specificity** towards the review, as they do not address the issues raised by the reviewer. For example, statements such as "your experience did not meet your expectations" may refer to all kinds of problems. TripAdvisor explicitly encourages response writers to address the problems mentioned in the reviews [Barsky and Frame, 2009]. We also assume that the businesses appear more concerned for their guests when the responses are more detailed. Zhang and Vásquez [2014] note that highly general responses are likely created by using a template or copy-pasting from other responses and without the response creators taking note of the details in the review. Considering this strategy, they question whether the businesses indeed use the feedback to improve their services as they often claim, as in example 13. This doubt highlights that poorly crafted responses are not necessarily helpful in improving the eWOM of a business.

(13) **Generated response:** Many thanks for taking the time to write your review [...]. We are disappointed to learn that your experience did not meet your expectations. We will certainly use your feedback to further improve our offering for future diners.

Other generated responses such as example 14b are more specific and **copy problems** mentioned by the reviewers. However, after referencing the issue, they **do not explain it or put it into perspective** which is normally done in human-written responses. In example 14c, the human response writer justifies the food prices of the

restaurant with the food quality ("fresh, local produce") and the restaurant ambiance ("fine dining experience and idyllic location").

(14) a. **Review:** Too expensive for the food they offer and for the service [...]
b. **Generated response:** [...] I am sorry to hear that you found our food and service to be too expensive. I hope you will visit us again soon.
c. **Human-written response:** [...] We are sorry you feel that Westbeach was too expensive. We pride ourselves on offer fresh, local produce and with the fine dining experience and idyllic location our customers are generally very happy with their meal. [...]

Levy et al. [2013] found that explaining the reasons for issues is one of the most typical moves made in responses of "high-performing" businesses. They thus consider it a good practice that sets a standard for other aspiring companies.

While it is advisable that responses address the issues raised in the reviews, the models sometimes **hallucinate problems** not mentioned in the reviews which is the third strategy of reacting to negative reviews. In example 15, the model accentuates the restaurant's goal to make their guests full, although the reviewer did not complain about portion sizes. We found that "leave full and happy" is an expression that occurred multiple times in the restaurant's training set. However, it usually appeared only in contexts where reviewers broached the issue of portion size.

(15) **Generated response:** [...] We work hard to make sure our guests leave *full* and happy, but we're sorry if you felt that wasn't the case this time [...]

We have seen in section 6.1.2 that copying parts from the reviews for the responses is a legitimate way of creating more specific responses. However, directly **copying long sequences such as entire sentences from the review** results in awkward formulations as in example 16b. Regarding content, apologising for the problem mentioned in the review is a good move. However, paraphrasing the issue would have made the response more natural-sounding.

(16) a. **Review:** [...] Also - be careful about the time of day when you make reservations. *The sun can shine directly in your eyes around late afternoon and there isn't much that the staff can do to mitigate this.* [...]
b. **Generated response:** [...] I am sorry that you did not enjoy the service you received and that *the sun can shine directly in your eyes around late afternoon and there isn't much that the staff can do to mitigate this.* [...]

In other cases, full sentences from the review are reproduced in the responses without addressing them further as in example 17b. Following best practices, customers' problems should not only be referenced but also explained, and the following steps of action to solve the problems should be named [Levy et al., 2013].

(17) a. **Review:** [...] They told me I had to use the toilet on the 1st floor because I didn't have a table which meant I have to [...] walk miles to the toilet. [...] *If you are planning to go its probably best to book a table but it does have a minimum spend.* Not the best.

b. **Generated response:** Thank you very much for taking your time to provide feedback. I am sorry that you had to use the toilet on the 1st floor. *If you are planning to go its probably best to book a table but it does have a minimum spend.* I hope you will be back again soon.

There were also cases where the models did **not react at all to criticism** of the customers. In example 18b, a negative review is answered with a response typical of positive reviews. It is likely that such a behaviour gives the impression that the response author did not read the review, resulting in a customer's feeling of not being taken care of.

(18) a. **Review:** [...] We generally enjoyed our meal until the bill came and we saw that the risotto course was £38 - for a medium sized plate of rice and vegetables (no meat). We were offered it as a special (but no price was mentioned). [...] This is poor practise. The waitstaff did not handle our complaint well.

b. **Generated response:** Thank you very much for taking the time to share your experience with us, we are pleased to know that you enjoyed your time with us. We look forward to see you at your next visit.

In this section, we have seen that the models apply various strategies of responding to negative reviews that diverge from those typically used by human response writers. Furthermore, the models fail to identify all of the bad reviews, and, consequently, do not react appropriately to them. In the next section, we approximate the models' recall of negative reviews to discover how many negative reviews remain undetected.

6.2.2 Identifying Negative Reviews

In contrast to other works [Gao et al., 2019a; Katsiuba et al., 2022; Kew et al., 2020], our models have no explicit information about customer satisfaction in form

of star ratings or sentiment scores. This necessitates a check as to whether the models can detect negative feelings from the reviews alone. In order to estimate the models' capabilities of sentiment analysis, we analysed all ground truth and generated responses for reviews with 1- or 2-star ratings. For each of these responses, we checked whether it contained any of the tokens or word stems[2] characteristic of typical reactions to negative reviews such as:

- apologising for problems and mistakes,
- acknowledging and referencing issues mentioned in customer feedback,
- promising that the problems will be solved and not re-occur, and
- inviting reviewers for follow-up communication [Zhang and Vásquez, 2014].

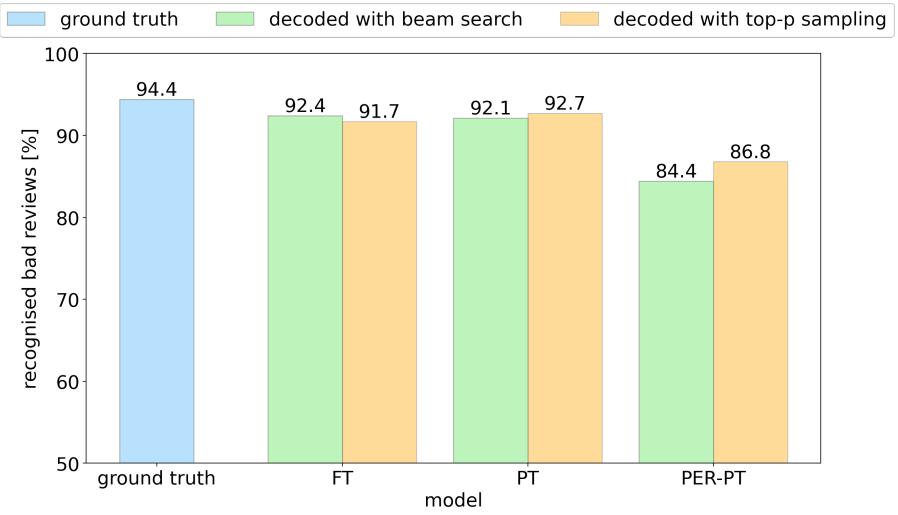


Figure 15: Percentage of responses for 1- or 2-star reviews containing at least one word characteristic of responses to negative reviews

Figure 15 shows the percentages of responses to negative reviews that contained any of the words from the set of negative review response indicators. On the left, we see that 94.4% of the ground truth responses to reviews with 1- or 2-star ratings contain at least one of the indicative words for complaint responses.

Inspecting the remaining results on figure 15, we find that the trained models perform differently in identifying negative reviews and reacting accordingly: The FT and PT models address almost as many negative reviews as the ground truth responses, while the PER-PT model is weaker at detecting negative reviews. Top-p

[2]We compiled this set of tokens and word stems characteristic of complaint responses: {"sorry", "disapp", "sad", "apolog", "upset", "unhappy", "complain", "regret", "unfortunate", "email", "contact", "in touch", "investigate", "dishearten", "negative", "unsatisfactory", "issue", "concern"}.

sampling improves the PER-PT model's sentiment analysis abilities.[3] One possible reason for this is that the PER-PT model outputs only few different responses for certain restaurants when using beam search which are exclusively aimed at positive reviews (cf. section 6.1.1). As beam search is a maximisation-based decoding strategy, these few responses represent the most probable ways for the model of reacting to reviews. With top-p sampling, the model generates less likely (but more appropriate) responses that acknowledge the criticism.

Overall, the recall of negative reviews is high for all trained models and decoding strategies: All models reacted appropriately to negative reviews (according to the word list) in at least 89% of those cases where the ground truth responses showed a behaviour that is typical of responses to negative reviews (as we can deduce from figure 15). Therefore, we conclude that the models have learned to recognise negative reviews solely by their texts relatively well and to apply one or more of the typical strategies for responding to negative reviews.

6.3 Hallucination

Large neural language models are known to produce natural-sounding text which is why they have been used to create conversational agents. While texts generated by such agents are perceived as highly engaging [See et al., 2019], they often contain claims for which there is no evidence in the inputs. This phenomenon has been referred to as "hallucinating" [Dziri et al., 2022; Kew and Volk, 2022]. Hallucinating unaccounted for facts is problematic, as the model runs risk of spreading misinformation [Dziri et al., 2022].

Manual inspection has shown that our models also make claims that are not based on evidence in the input reviews. In the following, we will investigate two kinds of topical hallucination, i.e. hallucination of restaurant names and cuisine types.

For inspecting hallucinated restaurant names, we counted the occurrences where the models produced a restaurant name, although it did not appear in the input review. We distinguished between "correct" mentions of the restaurant which the input review referred to and "incorrect" mentions of other restaurants.

To analyse which models tend to hallucinate cuisines, we searched for a set of cui-

[3] A concrete example of where top-p sampling helped to detect a negative review while beam search failed to do so can be found in section 6.1.3.

sine names[4] and checked whether they only occurred in the response but not in the review. Same as for the restaurant names, we differentiated between mentions of cuisines that suited the reviewed restaurants (based on their classification by TripAdvisor) and mentions that did not.

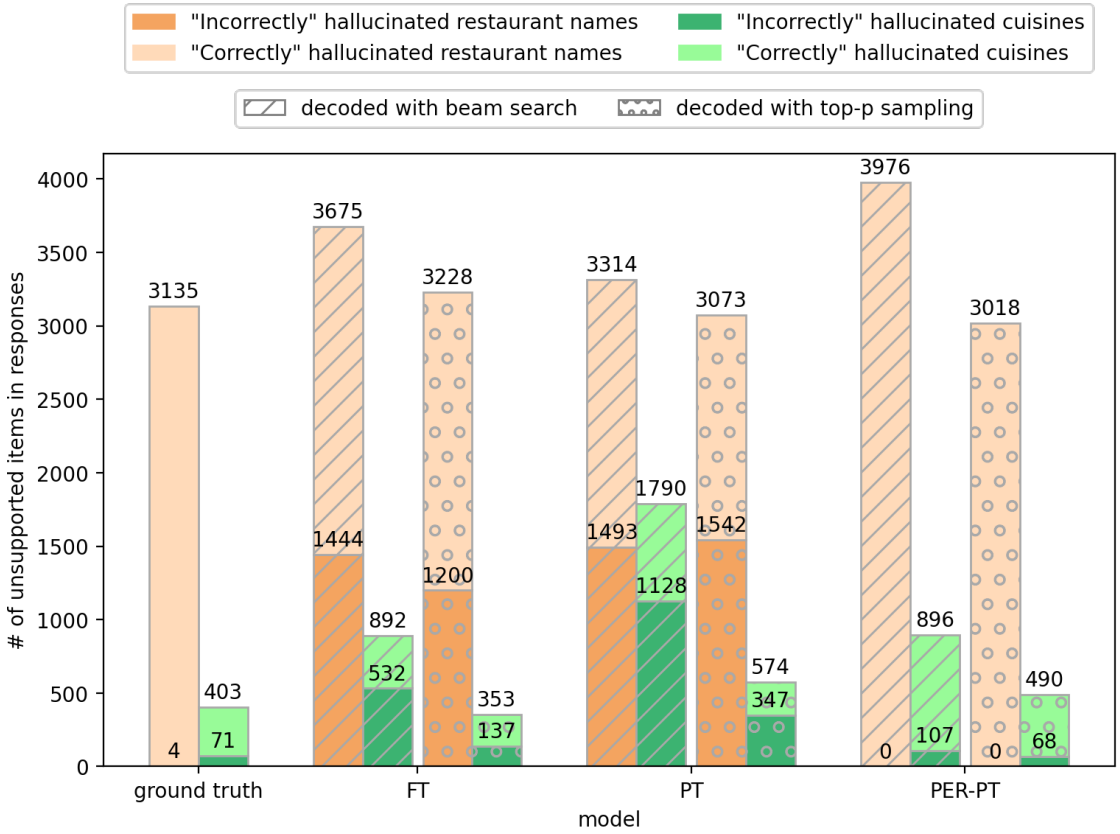


Figure 16: Occurrences of restaurant names and cuisines in the responses that did **not** occur in the corresponding review

Restaurant Names

Restaurant names occur frequently in some restaurants' review responses as shown in figure 8.e) (page 30). Therefore, it is unsurprising that the models include restaurant names in many of the generated responses as well. Figure 16 shows that the ground truth responses contain numerous mentions of restaurant names that have not been used in the reviews. Since these responses have been written by humans employed by specific restaurants, the claim of writing in the name of a restaurant without evidence in the reviews is not problematic. The same goes for the PER-PT model:

[4]These are the cuisines we searched for: {"Irish", "Mexican", "Peruvian", "Thai", "Italian", "Indian", "French", "German", "British", "American", "Spanish", "Steakhouse", "Asian", "Caribbean", "Mediterranean", "Seafood"}. Food from these cuisines is offered in our data set's restaurants.

As we trained it with individual restaurants' data sets, it only expects reviews for single restaurants. Thus, it can safely respond in the name of the restaurant towards which it was personalised. The FT and PT models, on the other hand, were trained on many restaurants' data sets. Therefore, they cannot know which restaurant the review was intended for unless they can deduce this from the review text.

Figure 16 shows that all models mention restaurant names more frequently than the ground truth responses when using beam search. This indicates that the models have learned that including restaurant names is highly characteristic of review responses. The generated responses of the PT and PER-PT models mention restaurant names similarly often as the ground truth responses when using top-p sampling.

Naming other restaurants' names is a behaviour that we observe almost exclusively for the FT and PT models. We assume that in the majority of these cases, the models write responses in the name of restaurants that have nothing to do with the input review, which is an inappropriate behaviour. The PER-PT model never mentioned restaurant names for which it did not author responses. The ground truth responses have four mentions of restaurants that differ from the reviewed restaurants. They refer to restaurants that the response writers recommend to the reviewers as in example 19. These instances are unproblematic.

(19) **Excerpt from response of "Paternoster Chop House:"** I hope you will visit us again, although it seems you would prefer somewhere more fine dining. Perhaps try [...] *Coq d'Argent.*

The FT and PT models infer the "correct" restaurants from the input reviews in more than half of all cases of hallucination. We hypothesise that cues such as location information help the models to deduce the respective restaurant, such as in example 20. Both reviews are for "Founding Farmers", the only restaurant in the data set located in Washington D.C., and both contain indications to the restaurant's location (marked in italics). We assume that the model used this information to generate a typical response of the restaurant that can be found in its training set.

(20) a. **Review 1:** [...] We make this a must on every visit to *DC.* [...]
b. **Review 2:** Delicious, fresh ingredients, and great energy in this local restaurant *near George Washington U.* We stopped here for dinner as it was in walking distance from our hotel and in *GWU vicinity.* [...]
c. **Generated response to reviews 1 and 2:** [...] I am thrilled to read how much you enjoyed your trip to *Founding Farmers*! [...]

Cuisines

Figure 16 shows that mentioning cuisines that do not appear in the reviews is more common for the generated than for the ground truth responses. These mentions occurred more often when using beam search rather than top-p sampling.

We can find similar patterns as for the restaurant names w.r.t. naming cuisines that are not characteristic of the reviewed restaurants: The PER-PT model, which is personalised towards single restaurants, infrequently mentions non-characteristic cuisines. The ground truth responses also have only infrequent mentions of cuisines that were not part of the reviews and that are not the main cuisine of the restaurant. These cases occur, for example, when the responses mention specially themed festivals as in example 21 or dishes and drinks with a cuisine type in their names as in example 22.

(21) **Seafood restaurant's response:** [...] We look forward to welcoming you back again soon, perhaps during our *Mediterranean* food festival [...].

(22) **American restaurant's response:** I hope you had the chance to try the newest drink we are featuring, *Irish* Mule! It's a huge hit!

The PT and the FT models produced many more mentions than the PER-PT and the ground truth responses of cuisines that are not characteristic of the reviewed restaurant. This gives rise to the assumption that most of these mentions are inappropriate in their context. Example 23 illustrates how the FT model generated a response for a Thai restaurant, even though the review did not have any indications of which kind of cuisine the restaurant is specialised in. The generated response is highly inappropriate in this context since the review originally referred to an American restaurant.

(23) a. **Review:** Excellent food, very reasonable prices, good service at the bar, pleasant discovery. Would go back often.

b. **Generated response:** [...] I am very happy to hear that you have enjoyed your *Thai experience* with us [...]

In summary, all models as well as the ground truth responses mention restaurant names and cuisines that have not been part of the input reviews. However, claiming to write in the name of the wrong restaurant or naming cuisines that are unrelated to the input review are behaviours that we observed frequently for the PT and the

FT models but rarely for the PER-PT model and the ground truth responses. Given that the hallucinations are correct guesses, they are unproblematic. Wrong claims, however, lead to inappropriate responses. With a growing number of businesses in the data set, it becomes less likely that the guesses made by the models are acceptable. Personalised response generation can help to prevent wrong claims in this scenario.

6.4 Evaluating Personalisation: Difficulties and Pitfalls

As one goal of this thesis is to evaluate the performance of prefix-tuning at generating personalised review responses, we need to quantify the success at personalisation. However, the task of measuring the degree of personalisation in a generated text is far from trivial. Various metrics which find use in other text generation tasks such as translation or summarisation have been reported in studies on personalisation, but their usage in this context has been debated and argued against. Some other metrics have the exclusive purpose of measuring the degree of personalisation, but their expressiveness and correlation with human judgement have not been thoroughly analysed. In this section, we revisit some of the metrics presented in section 2.3 and discuss potential issues of applying them to evaluation of personalisation.

6.4.1 Reference-Based Metrics

Of all metrics in the personalisation studies described in section 2.2, BLEU [Papineni et al., 2002] is the most popular. However, its efficacy in this task is debatable. Liu et al. [2016] have shown that BLEU correlates weakly or not at all with human judgement (depending on the data set) when evaluating dialogue response generation. Furthermore, Kew and Volk [2022] and Zhang et al. [2019] noted that using a limited number of references fails to account for the infinite possibilities of how one can respond to an utterance. This makes BLEU a less than ideal metric on open-ended tasks. Example 24 shows that both the ground truth as well as the generated response are legitimate in a given context. However, the BLEU score for the system's response would be 0, i.e. it would be rated maximally poorly, although the text is appropriate given the review.

(24) a. **Review:** [...] the food [for this Hard Rock] is delicious (as all the cafes around de US). The only thing that I dont like is when is someones birthday and the waiter screams behind you haha.

b. **Ground truth response:** We do have fun with birthdays here at the Hard Rock!

c. **Generated response:** Thank you for visiting our Hard Rock Café Orlando as well as our other locations!

Having more references increases the chances that different valid responses are rated with a decent BLEU score. However, collecting more references is costly, and covering all possible responses to a review is still unfeasible.

ROUGE [Lin, 2004] and unigram F1 [Wang et al., 2021] are less frequently applied as evaluation metrics in the area of personalisation, but their potential problems are similar to the ones of BLEU. ROUGE has very low correlation with human judgement in evaluating response generation [Liu et al., 2016], and both metrics work with a limited number of references that cannot capture all possible responses as well. An additional property of F1 is that it only operates on the level of unigrams. Since many stylistic characteristics span across several tokens, considering only unigrams fails to acknowledge sequences of tokens that are characteristic of a persona.

Another kind of reference-based metric used in the previous personalisation studies is based on embeddings [Ma et al., 2021; Zhang et al., 2019]. A drawback of such metrics is that they are not intended for texts longer than one sentence. Since review responses are usually composed of several sentences, it is unclear whether using such embedding-based metrics is appropriate. Liu et al. [2016] also found that different embedding-based metrics have low or no correlation with human judgement in the evaluation of the dialogue response generation task.

6.4.2 Personalisation Metrics

The most prominent concern regarding the personalisation metrics listed in section 2.3.3 is that they have been used in only few studies. This differentiates them from the reference-based metrics listed in 2.3.1 which are far more prominent in NLP research. Consequently, in most cases, their usefulness has only been assumed but not closely scrutinised. Future research is therefore still necessary in order to investigate the effectiveness of these personalisation metrics. In this section, we note some metric-specific problems we detected while applying them to our task.

Persona-R/-P/-F1 [Lv et al., 2020; Ma et al., 2021] fails to capture stylistic characteristics spanning across multiple tokens, as it operates only on the unigram level.

P-Cover [Ma et al., 2021; Song et al., 2019] and PTSal [Miyazaki et al., 2021] weight

words to calculate a response's score. The underlying idea behind this is that not all tokens reveal to the same extent whether a text is characteristic of a persona. The goal of P-Cover is to reward those responses with high scores which contain words that are infrequent in a personalisation corpus. PTSal does the same for responses with words that are considered characteristic of a persona. Responses that contain only common words, on the other hand, have a low degree of personalisation and are thus punished with low scores. What sounds plausible at first, however, is implemented in a way that subverts the expressiveness of these weights: Both P-Cover and PTSal calculate their overall score by averaging over the words' weights.[5] If most of the considered words have low weights and a few have high weights, the reward of the higher weighted words is cancelled out by the lower weights of the less important words. This results in weak scores for potentially highly personalised responses which is undesirable for a personalisation metric. The normalisation procedure is illustrated in an example for the metric P-Cover in table 8 (appendix A).

With Per-Hits@k [Wang et al., 2021], one does not directly compare the generated texts with the training set responses but uses the training sets to train as many language models as personae in the data set. Wang et al. [2021] demonstrate that the results of Per-Hits@k are highly correlating when using two different language model architectures to calculate the perplexity values, thus partially alleviating the concern that perplexity is highly variable across different language model architectures [Jin et al., 2022]. The other risks of language models described in section 2.3.2, however, still remain. A further drawback of this evaluation metric is its computational cost, as it requires a user to train as many language models as personae. Also, the correlation between the Per-Hist@k scores and human judgement has not been subject of analysis yet, thus the usefulness of this metric is still questionable.

Overall, the evaluation of personalisation poses a challenge. We have seen in section 2.2.1 that different definitions of the concept "style" exist which all remain vague to some extent. This complicates defining a gold standard for personalisation evaluation. Due to the lack of a standard evaluation procedure, different researchers use existing metrics from other text generation tasks or develop their own metrics. However, these have not yet undergone the scientific scrutiny that would be necessary to evaluate their expressiveness. Furthermore, we argue that the open-endedness of response generation will always restrict the usefulness of automatic metrics.

[5]The averaging for P-Cover happens in equation 2.2 (page 14) where the mean ITF score of the words in the overlap set of the training set response and the generated response is calculated. The overall PTSal score of a generated response is acquired by calculating the PTSal for each word in the response according to equation 2.5 (page 15) and averaging over these PTSal values.

7 Conclusion

In this thesis, we applied prefix-tuning [Li and Liang, 2021] for adapting the pre-trained model BART [Lewis et al., 2020] towards review response generation. We investigated to what extent a prefix-tuned model approximated the performance of a fine-tuned model with regard to general review response generation as our first RQ. Furthermore, we leveraged the modularity offered by prefixes to tackle the research gap of personalised review response generation. In doing so, our aim was to answer RQ2, i.e. how well tuned prefixes could reproduce the stylistic characteristics of the data set's individual restaurants in personalised review response generation. For RQ3 we examined human-written reviews and responses as well as generated responses in order to identify the largest challenges of review response generation.

To answer these questions, we compiled our own data set based on detailed analysis of previous data sets used in personalisation studies. We explained which considerations we took into account for creating our data set, and we described the characteristic properties of our data set.

With respect to RQ1, we have shown in our experiments that prefix-tuning is indeed a viable alternative to fine-tuning for adapting a pre-trained model to the downstream task of review response generation, especially in scenarios where memory efficiency is crucial. We demonstrated that the prefix-tuned model approximates the fine-tuned model's performance in all aspects except for inter-textual diversity.

Regarding RQ2, we have shown that prefix-tuning offers a lightweight solution for producing review responses with a higher degree of personalisation than conventional fine-tuning.

While answering the first two RQs, we found that the inter-textual diversity of the generated responses was unsatisfactorily low when using the de facto decoding strategy of beam search. Using the sampling-based decoding strategy of top-p sampling [Holtzman et al., 2019] increased the diversity across the responses to a large extent. It also consistently helped improving the responses' specificity. In the personalised responses, the restaurant's characteristics were present to a similar degree as in the human-written responses when using top-p sampling. These findings strongly sug-

gest that top-p sampling is more suitable for this type of open-ended text generation task than beam search.

Considering RQ3, we identified some common challenges of review response generation: hallucination, lack of specificity, and evaluation of personalisation.

Hallucinated content, i.e. content unaccounted for by the input review, was mainly produced by the models trained within the scope of RQ1. We found that the problem of hallucinating could be greatly alleviated by training personalised prefixes on smaller restaurant-specific data sets. However, this comes at the cost of sacrificing the diversity of the training data. One promising possibility of combining all the benefits of prefix-tuning on personal data sets and generalising knowledge from diverse data sets would be to first fine-tune a pre-trained language model on a larger corpus of review-response pairs of different restaurants before tuning personalised prefixes for individual businesses with the fine-tuned model. Future work is required to evaluate the results of this training approach.

Another common problem of response generation systems including review response generation systems is their tendency to produce highly universal responses [Kew and Volk, 2022; Li et al., 2016a] that are unspecific to the input. As mentioned above, top-p sampling improved the responses' specificity. However, we did not take an empirical approach to find the optimal hyperparameters, nor did we try similar sampling-based strategies. Thus, future work is needed in order to analyse the effect of changing the hyperparameter settings of top-p sampling or using other sampling-based strategies. An alternative approach for generating more specific reviews proposed by Kew and Volk [2022] is to use a curated training set with the most generic responses filtered out. Combining training set filtering with prefix-tuning as the learning strategy in order to generate specific and personalised responses yields further promising experiments for future work.

Future research is also needed to investigate the personalisation metrics that we described and used in this thesis. To date, scrutinising the congruence of automatic personalisation metrics and human judgement is a research gap. Filling this gap would be an important step towards establishing an evaluation standard for personalisation which is not given in today's personalisation evaluation jungle. We believe that our findings and observations serve as a solid basis on which future work may be built.

References

A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, 2018.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*, San Diego, California, USA, 2015.

J. Barsky and C. Frame. Handling online reviews: Best practices, 2009. URL `https://cdn.tripadvisor.com/pdfs/ExpertTips_HandlingOnlineReviews.pdf`. (last accessed 22 September 2022).

N. L. Chan and B. D. Guillet. Investigation of social media marketing: How does the hotel industry in Hong Kong perform in marketing on social media websites? *Journal of Travel & Tourism Marketing*, 28(4):345–368, 2011.

V. Creelman. Sheer outrage: Negotiating customer dissatisfaction and interaction in the blogosphere. In E. Darics, editor, *Digital Business Discourse*, pages 160–185. Palgrave Macmillan, London, England, 2015.

S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, Online, 2020.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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, USA, 2019.

N. Dziri, H. Rashkin, T. Linzen, and D. Reitter. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083, 2022.

A. Fan, M. Lewis, and Y. Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, 2018.

C. Gao, J. Zeng, X. Xia, D. Lo, M. R. Lyu, and I. King. Automating app review response generation. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 163–175, San Diego, California, USA, 2019a.

J. Gao, M. Galley, and L. Li. Neural approaches to conversational AI. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298, 2019b.

A. Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly, Sebastopol, California, USA, 2nd edition, 2019.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, Massachusetts, USA, 2016. `http://www.deeplearningbook.org`.

J. Holmes. *An Introduction to Sociolinguistics*. Routledge, London, England, 4th edition, 2013.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text *de*generation, 2019. URL `https://arxiv.org/abs/1904.09751`.

D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205, 2022.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, 2017.

D. Katsiuba, T. Kew, M. Dolata, and G. Schwabe. Supporting online customer feedback management with automatic review response generation. In *The 55th Hawaii International Conference on System Sciences*, pages 226–236, Hawaii, USA, 2022.

T. Kew and M. Volk. Improving specificity in review response generation with data-driven data filtering. In *Proceedings of The Fifth Workshop on e-Commerce and NLP*, pages 121–133, Dublin, Ireland, 2022.

T. Kew, M. Amsler, and S. Ebling. Benchmarking automated review response generation for the hospitality domain. In *Proceedings of Workshop on Natural Language Processing in E-Commerce*, pages 43–52, Barcelona, Spain, 2020.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526, 2017.

S. E. Levy, W. Duan, and S. Boo. An analysis of one-star online reviews and responses in the Washington, D.C., lodging market. *Cornell Hospitality Quarterly*, 54(1):49–63, 2013.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, USA, 2016a.

J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, 2016b.

J. Li, R. Jia, H. He, and P. Liang. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana, USA, 2018.

X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online, 2021.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004.

C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, USA, 2016.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2017. URL `https://arxiv.org/abs/1711.05101`.

P. Lv, S. Feng, D. Wang, Y. Zhang, and G. Yu. PersonaGAN: Personalized response generation via generative adversarial networks. In *Database Systems for Advanced Applications: 25th International Conference, Proceedings, Part I*, page 570–586, Jeju, South Korea, 2020.

Z. Ma, Z. Dou, Y. Zhu, H. Zhong, and J.-R. Wen. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Online, 2021.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England, 2008.

P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium, 2018.

C. Miyazaki, S. Kanno, M. Yoda, J. Ono, and H. Wakaki. Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 178–189, Singapore and Online, 2021.

A. Napolitano. Image repair or self-destruction? A genre and corpus-assisted discourse analysis of restaurants' responses to online complaints. *Critical Approaches to Discourse Analysis across Disciplines*, 10(1):135–153, 2018.

I. S. Pantelidis. Electronic meal experience: A content analysis of online restaurant comments. *Cornell Hospitality Quarterly*, 51(4):483–491, 2010.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002.

A. A. Parikh, C. Behnke, B. Almanza, D. Nelson, and M. Vorvoreanu. Comparative content analysis of professional, semi-professional, and user-generated restaurant reviews. *Journal of Foodservice Business Research*, 20 (5):497–511, 2017.

J. Phang, T. Févry, and S. R. Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks, 2018. URL `https://arxiv.org/abs/1811.01088`.

S. T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, 2014.

M. Popović. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, 2015.

D. Proserpio and G. Zervas. Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science*, 36 (5):645–665, 2017.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018. URL `https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf`. (last accessed 13 May 2022).

A. See, S. Roller, D. Kiela, and J. Weston. What makes a good conversation? How controllable attributes affect human judgments. 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 1702–1723, Minneapolis, Minnesota, USA, 2019.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016.

K. Shuster, S. Humeau, A. Bordes, and J. Weston. Image-Chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online, 2020.

E. M. Smith, D. Gonzalez-Rico, E. Dinan, and Y.-L. Boureau. Controlling style in generated dialogue, 2020. URL `https://arxiv.org/abs/2009.10855`.

H. Song, W.-N. Zhang, Y. Cui, D. Wang, and T. Liu. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5190–5196, Macao, China, 2019.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, USA, 2015.

B. A. Sparks, K. K. F. So, and G. L. Bradley. Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. *Tourism Management*, 53:74–85, 2016.

F.-G. Su, A. R. Hsu, Y.-L. Tuan, and H.-Y. Lee. Personalized dialogue response generation learned from monologues. In *Interspeech*, pages 4160–4164, Graz, Austria, 2019.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, page 3104–3112, Cambridge, Massachusetts, USA, 2014.

J. Tiedemann. News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing*, 5: 237–248, 2009.

D. Wang, N. Jojic, C. Brockett, and E. Nyberg. Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150, Copenhagen, Denmark, 2017.

Z. Wang, L. Luo, and D. Yang. Personalized response generation with tensor factorization. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics*, pages 47–57, Online, 2021.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020.

Y. Wu, X. Ma, and D. Yang. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online, 2021.

K. L. Xie, Z. Zhang, Z. Zhang, A. Singh, and S. K. Lee. Effects of managerial response on consumer eWOM and hotel performance: Evidence from TripAdvisor. *International Journal of Contemporary Hospitality Management*, 28(9):2013–2034, 2016.

Z. Xu, N. Jiang, B. Liu, W. Rong, B. Wu, B. Wang, Z. Wang, and X. Wang. LSDSCC: A large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2070–2080, New Orleans, Louisiana, USA, 2018.

M. Yang, W. Tu, Q. Qu, Z. Zhao, X. Chen, and J. Zhu. Personalized response generation by dual-learning based domain adaptation. *Neural Networks*, 103(C): 72–82, 2018.

S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, 2018.

W.-N. Zhang, Q. Zhu, Y. Wang, Y. Zhao, and T. Liu. Neural personalized response generation as domain adaptation. *World Wide Web*, 22(4):1427–1446, 2019.

Y. Zhang and C. Vásquez. Hotels' responses to online reviews: Managing consumer dissatisfaction. *Discourse, Context & Media*, 6:54–64, 2014.

Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. DialoGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics: System Demonstrations, pages 270–278, Online, 2020.

L. Zhao, K. Song, C. Sun, Q. Zhang, X. Huang, and X. Liu. Review response generation in E-Commerce platforms with external product information. In *The World Wide Web Conference*, page 2425–2435, New York, New York, USA, 2019.

Y. Zheng, G. Chen, M. Huang, S. Liu, and X. Zhu. Personalized dialogue generation with diversified traits, 2019. URL `https://arxiv.org/abs/1901.09672`.

Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1097–1100, New York, New York, USA, 2018.

Curriculum Vitae

Personal Details

Helen Schaller
Rütiwisstrasse 12
8604 Volketswil
helen.schaller@uzh.ch

Education

2016-2020	Bachelor of English and German Language and Literature at University of Zurich
2020-2022	Master of Computational Linguistics and Language Technology at University of Zurich

Professional and Part-Time Activities

2016-2021	Various part-time jobs with different employers
Spring semester 2022	Tutoring for Fundamentals of Speech Sciences and Signal Processing

A Tables

Restaurant	Cuisine	Place	Price category	RR Pairs
Hard Rock Cafe	American	Orlando, USA	“mid-range”	5,854
Margaritaville Las Vegas	Mexican	Las Vegas, USA	“mid-range”	3,953
Jimmy Buffett’s Margaritaville	American	Orlando, USA	“mid-range”	3,502
Top of the World	American	Las Vegas, USA	“fine dining”	3,033
Hash House a Go Go	American	Las Vegas, USA	“mid-range”	2,945
The Old Storehouse Bar Restaurant	Irish	Dublin, Ireland	“mid-range”	2,847
The Savoy Grill	British	London, England	“fine dining”	2,821
Pollen Street Social	British	London, England	“fine dining”	2,700
Founding Farmers	American	Washington DC, USA	“mid-range”	2,700
Salt House	Contemporary	Cairns, Australia	“mid-range”	2,472
The Boxty House	Irish	Dublin, Ireland	“mid-range”	2,409
Senor Frog’s Las Vegas	Mexican	Las Vegas, USA	“mid-range”	2,401
The Smugglers Cove	Bar	Liverpool, England	“mid-range”	2,329
Fazenda Rodizio Bar and Grill	Steakhouse	Liverpool, England	“mid-range”	2,316
Cinnamon Club	Indian	London, England	“fine dining”	2,254
Eiffel Tower Restaurant at Paris Las Vegas	French	Las Vegas, USA	“fine dining”	2,188
Quaglino’s	British	London, England	“fine dining”	2,138
Citrique Restaurant	Seafood	Surfers Paradise, Australia	“fine dining”	2,038
Paternoster Chop House	British	London, England	“mid-range”	1,984
Havana 1957 Cuban Cuisine Espanola Way	Caribbean	Miami Beach, USA	“mid-range”	1,976
Toro Toro	Steakhouse	Miami, USA	“mid-range”	1,952
Donatello	Italian	Brighton, England	“mid-range”	1,889
Andiamo Italian Steakhouse	Italian	Las Vegas, USA	“fine dining”	1,863
Browns Bar Brasserie	Bar	Liverpool, England	“mid-range”	1,827
Madison	Steakhouse	London, England	“mid-range”	1,827
100 Wardour St	British	London, England	“mid-range”	1,778
The Oliver Plunkett	Irish	Cork, Ireland	“mid-range”	1,735

Restaurant	Cuisine	Place	Price category	RR Pairs
Urban Tandoor	Indian	Bristol, England	“mid-range”	1,704
La Tasca	Spanish	Liverpool, England	“mid-range”	1,682
The Cowfish Sushi Burger Bar	Asian	Orlando, USA	“mid-range”	1,605
Punjab	Indian	London, England	“mid-range”	1,603
Fire Restaurant Lounge	Steakhouse	Dublin, Ireland	“fine dining”	1,589
Cinnamon Kitchen City	Indian	London, England	“mid-range”	1,558
Quinlans Seafood Bar	Irish	Killarney, Ireland	“mid-range”	1,511
German Gymnasium	German	London, England	“mid-range”	1,481
The Rajdoot	Indian	London, England	“mid-range”	1,456
Carpathia Rooftop Bar Restaurant	Bar	Liverpool, England	“mid-range”	1,456
Revolution Albert Dock	Bar	Liverpool, England	“mid-range”	1,452
Tozi Restaurant Bar	Italian	London, England	“mid-range”	1,399
Red Torch Ginger	Asian	Dublin, Ireland	“mid-range”	1,369
Coq d’Argent	French	London, England	“fine dining”	1,333
Maha	Mediterranean	Melbourne, Australia	“fine dining”	1,314
Santorini by Georgios	Seafood	Miami Beach, USA	“mid-range”	1,309
Galvin La Chapelle	French	London, England	“fine dining”	1,305
Bistrot Pierre	French	Leicester, England	“mid-range”	1,264
Dundee’s Restaurant on the Waterfront	Seafood	Cairns, Australia	“mid-range”	1,257
Leicester Square Kitchen	Mexican	London, England	“fine dining”	1,233
WestBeach	Seafood	Bournemouth, England	“mid-range”	1,212
Chaophraya Liverpool	Asian	Liverpool, England	“mid-range”	1,196
Pizza Pilgrims	Italian	London, England	“cheap eats”	1,193

Table 7: Restaurants in our data set

ITF	Word
0.5	margaritas
0.15	thank
0.08	you
0.07	for

Training set: Thank you for this review! We pride ourselves in making fantastic margaritas !

Generated 1: Great to hear the margaritas were to your liking.
→ weighted personalisation score: **0.5**

Generated 2: Thank you for the 5*. Great to hear the margaritas were to your liking.
→ weighted personalisation score: $\frac{(0.5+0.15+0.08+0.07)}{4} =$ **0.2**

Table 8: Exemplary calculation of weighted personalisation metric P-Cover. (Left) Lookup table with weights of words that occur in the overlap sets of the training set response and the generated responses. Large weights indicate that the words are less frequent. (Right) Made up examples of a training set response and two generated responses. The highlighted words in the generated sentences occur in the training set response as well. Although the second generated response is more similar to the training set response, it has a lower P-Cover score than the first generated response due to the averaging process.

Table 8 illustrates why the weighting as applied in the personalisation scores P-Cover [Ma et al., 2021; Song et al., 2019] and PTSal [Miyazaki et al., 2021] is problematic. The weighting procedure is shown for the metric P-Cover. In the exemplary score calculation, two different generated responses are compared with a training set response. Both generated responses contain the rare word "margaritas" which is highly characteristic of the persona's style. The second generated response is more similar to the training set response than the first, since it expresses gratitude to the reviewer which the training set response does as well. Despite the second generated response's greater similarity to the persona response, the first generated response receives a higher P-Cover score, as the three low-weight tokens "thank", "you" and "for" from the second generated response shrink its score considerably by being included in the averaging process. In other words, the response with less overlap with the training set response is awarded a better score than the more similar generated response. This example illustrates how P-Cover results in counterintuitive scores that are not useful for evaluating personalisation.