



**Universität  
Zürich** <sup>UZH</sup>

Master's Thesis  
Presented to the Faculty of Arts and Social Sciences  
of the University of Zurich for the Degree of  
**Master of Arts**

# Attention Biasing for Knowledge-enhanced Review Response Generation

**Author: Sooyeon Cho**

Student ID Nr.: 21-737-085

Supervisor: Tannon Kew

Examiner: Prof. Dr. Martin Volk

Department of Computational Linguistics

Submission Date: 27.05.2023

## **Abstract**

Knowledge-enhanced review response generation, a task in the domain of Natural Language Generation (NLG), aims to produce contextually relevant responses to customer reviews by leveraging specific knowledge sources. A significant challenge in this task is the tendency of models to generate generic, universally applicable responses, lacking in specificity.

In addressing this issue, this thesis investigates a technique to enhance the specificity and informativeness of such responses within the hotel domain. We employ the cross-attention biasing method presented by Hazarika et al. [2021], designed to improve informativeness in knowledge-enhanced dialog generation tasks. This method modifies the attention distribution of an encoder-decoder NLG model, adjusting the allocation of attention to different parts of the input. This approach, applied within a sequence-to-sequence transformer model, is explored under various experimental setups in our work.

The results demonstrate that attention biasing helps to enhance the informativeness of the generated review responses at a significant level in restricted experimental settings: namely, where the input length of the provided knowledge snippet is restricted and the bias is given to the knowledge part that conveys contextual and natural language descriptions.

This research contributes to the development of more effective automated review response generation, with potential implications for enhancing customer engagement and reputation management in the hotel industry.

---

## Zusammenfassung

Die wissensgestützte Generierung von Antworten auf Kundenrezensionen, eine Aufgabe im Bereich der Natural Language Generation (NLG), zielt darauf ab, kontextuell relevante Antworten auf Kundenrezensionen zu erstellen, indem spezifische Wissensquellen genutzt werden. Eine besondere Herausforderung bei dieser Aufgabe besteht in der Tendenz der Modelle, generische und allgemeingültige Antworten zu erzeugen, denen es an Spezifität fehlt.

Um dieses Problem anzugehen, wird in dieser Arbeit eine Methode zur Verbesserung der Spezifität und Informativität solcher Antworten im Hotelbereich untersucht. Wir verwenden die von Hazarika et al. [2021] vorgestellte Cross-Attention-Biasing-Methode, die darauf abzielt, die Informativität in wissensgestützten Dialoggenerierungsaufgaben zu verbessern. Diese Methode modifiziert die Verteilung von Attention eines Encoder-Decoder-NLG-Modells, indem die Zuweisung von Attention zu verschiedenen Teilen der Eingabe angepasst wird. Dieser Ansatz wird in einem Sequence-to-Sequence-Transformer-Modell angewendet und unter diversen experimentellen Aufbauten in dieser Arbeit untersucht.

Die Ergebnisse zeigen, dass das Attention-Biasing zu einer signifikanten Verbesserung der Informativität der generierten Antworten unter bestimmten experimentellen Aufbauten beiträgt: nämlich dann, wenn die Eingabelänge des bereitgestellten Wissensausschnitts begrenzt ist und der Bias auf den Wissensteil zugewiesen ist, der kontextuelle und natürlichsprachliche Beschreibungen liefert.

Diese Forschungsarbeit trägt zur Weiterentwicklung einer effektiveren automatischen Antwortgenerierung auf Kundenrezensionen bei, mit möglichen Auswirkungen auf die Verbesserung des Kundenengagements und des Reputationsmanagements in der Hotelbranche.

# Acknowledgements

I would like to express my deepest gratitude to everyone who has contributed immensely to the completion of this Master's thesis. Their continuous support, guidance, and inspiration have been invaluable.

First and foremost, I extend my heartfelt gratitude to Tannon for his invaluable advice, supervision, constructive feedback, and the provision of essential research materials throughout my work on this thesis. His enduring support and intellectual insights have not only enriched this work but also significantly contributed to my academic growth.

To Prof. Dr. Martin Volk and the entire Department of Computational Linguistics, I offer my sincere thanks. The knowledge and guidance you offered further fueled my passion for computational linguistics.

I also owe a lot to Prof. Dr. Hye Yang Shin, Prof. Dr. Byung Jin Choi, and Prof. Dr. Hye-Yeon Chung for keep inspiring me to embark on the journey of language and linguistics. Your kindness, expertise, and dedication have served as a model for me in my continued studies.

Most importantly, I want to thank my family and Chris for their unwavering support throughout my studies. I am deeply grateful for the love, encouragement, and belief you have shown in me.

This journey would not have been possible without the support, guidance, and encouragement of you all. Thank you.

# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgement</b>	<b>iii</b>
<b>Contents</b>	<b>iv</b>
<b>List of Figures</b>	<b>vii</b>
<b>List of Tables</b>	<b>viii</b>
<b>List of Acronyms</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Research Questions . . . . .	2
1.3 Thesis Structure . . . . .	4
<b>2 Background</b>	<b>5</b>
2.1 Natural Language Generation . . . . .	5
2.2 Review Response Generation . . . . .	6
2.3 Knowledge-enhanced Text Generation . . . . .	8
2.3.1 Controllable Knowledge-enhanced Generation . . . . .	10
2.4 Transformer . . . . .	11
2.4.1 Transformer Architecture . . . . .	11
2.4.2 Biasing Attention Module . . . . .	13
2.5 Cross-Attention Biasing in Knowledge-enhanced Generation . . . . .	14
2.6 Evaluation Metrics for Natural Language Generation . . . . .	16
2.6.1 Reference-based Metrics . . . . .	16
2.6.2 Reference-free Metrics . . . . .	18
2.6.3 Evaluation Metrics for Knowledge-enhanced Generation . . . . .	18
<b>3 Preliminary Experiment</b>	<b>19</b>
3.1 Experimental Setup . . . . .	19
3.1.1 Dataset . . . . .	19

3.1.2	Formatting the Input . . . . .	20
3.1.3	Fine-tuning and Inference . . . . .	22
3.2	Results . . . . .	23
3.2.1	Evaluating Informativeness . . . . .	23
3.2.2	Evaluating Fluency . . . . .	25
3.3	Conclusion . . . . .	26
<b>4</b>	<b>RQ1 Experiments</b>	<b>27</b>
4.1	Methodology . . . . .	27
4.1.1	Cross-Attention Biasing . . . . .	28
4.1.2	Data Filtering . . . . .	28
4.2	Experimental Setup . . . . .	32
4.2.1	Dataset . . . . .	32
4.2.2	Preprocessing and Input Formatting . . . . .	33
4.2.3	Fine-tuning and Inference . . . . .	34
4.2.4	Setup for Experiment 2 . . . . .	36
4.2.5	Evaluation . . . . .	36
4.3	Results . . . . .	37
4.3.1	Experiment 1: Fine-tuning with Original Data (OD) . . . . .	38
4.3.2	Experiment 2: Fine-tuning with Filtered Data (FD) . . . . .	40
4.3.3	Comparison between OD and FD models . . . . .	42
4.3.4	Discussion . . . . .	43
<b>5</b>	<b>RQ2 Experiment</b>	<b>45</b>
5.1	Motivation . . . . .	45
5.2	Experimental Setup . . . . .	46
5.3	Results . . . . .	46
5.3.1	Automatic Evaluation . . . . .	47
5.3.2	Comparative Analysis of Inference Settings . . . . .	50
5.3.3	Key Findings . . . . .	53
<b>6</b>	<b>Further Analysis and Discussion</b>	<b>54</b>
6.1	Decoding the Results: A Comprehensive Analysis of the Experimental Outcomes . . . . .	54
6.1.1	Interpreting the Improvement of Precision <sub>k</sub> in OD and FD Models . . . . .	54
6.1.2	Understanding the Role of Knowledge Input Length in Attention Biasing Efficacy . . . . .	56
6.1.3	Understanding the Differential Impact of Attention Biasing on the Amenities Profile . . . . .	57
6.2	Comparative Analysis with ChatGPT . . . . .	58

6.2.1	Evaluation Metrics . . . . .	59
6.2.2	Qualitative Analysis . . . . .	60
6.2.2.1	Description Profile Analysis . . . . .	60
6.2.2.2	Amenities Profile Analysis . . . . .	62
6.2.2.3	Knowledge Profile Analysis . . . . .	64
6.2.3	Implications . . . . .	66
6.3	Limitations and Future Directions . . . . .	66
<b>7</b>	<b>Conclusion</b>	<b>69</b>
	<b>References</b>	<b>71</b>
<b>A</b>	<b>Tables</b>	<b>76</b>

# List of Figures

1	Knowledge-enhanced Text Generation . . . . .	9
2	Transformer Architecture . . . . .	12
3	Attention Biasing Knob . . . . .	15
4	Token Length Distribution of Topical Chat Dataset . . . . .	21
5	Visualization of Cross-Attention Biasing Profiles . . . . .	23
6	Token Length Distribution of the Hotel Dataset . . . . .	35
7	Recall and Effect Size of the Different Inference Settings . . . . .	51
8	Perplexity and Self-BLEU of the Different Inference Settings . . . . .	52
9	Average Output Length of the Models . . . . .	55



# List of Tables

1	Overview of the Topical Chat Dataset . . . . .	22
2	Informativess Evaluation for the Preliminary Experiment . . . . .	24
3	Perplexity Scores for the Preliminary Experiment . . . . .	25
4	Perplexity Score Distribution of Training Data . . . . .	31
5	Results for Experiment 1 (OD) . . . . .	38
6	Results for Experiment 2 (FD) . . . . .	41
7	Score Differences of OD and FD . . . . .	43
8	Results for Experiment 3 (KB100) . . . . .	47
9	Results of ChatGPT . . . . .	59
10	Standard Deviation for Experiment 1 (OD) . . . . .	76
11	Standard Deviation for Experiment 2 (FD) . . . . .	77
12	Results for Experiment 3 (KB40) . . . . .	78
13	Standard Deviation for Experiment 3 (KB100) . . . . .	79
14	Results for Experiment 3 (KB160) . . . . .	80

# List of Acronyms

BOS	Beginning of Sentence
CNN	Convolutional Neural Network
EOS	End of Sentence
FD	Filtered Data (Model)
IR	Information Retrieval
KB	Knowledge Bucket (Inference Setting)
NLG	Natural Language Generation
OD	Original Data (Model)
PPL	Perplexity
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
RQ	Research Question
Seq2seq	Sequence-to-Sequence
TF-IDF	Term Frequency—Inverse Document Frequency

# 1 Introduction

## 1.1 Motivation

Knowledge-enhanced review response generation is designed to automatically produce contextually relevant responses to customer reviews. It leverages specific knowledge sources to enhance the quality and relevance of the output. This technology, when sufficiently developed for full deployment, holds great promise for businesses in the hospitality domain, as it could substantially improve the way businesses interact with their customers and manage their online presence.

In today's digital age, managing online reviews has become a priority for businesses like hotels and restaurants. Effective responses to customer reviews can attract new customers, foster loyalty, enhance ratings, and increase sales (Zhao et al. [2019]; Cantalops and Salvi [2014]; Hassan et al. [2018]). However, the sheer volume of online reviews poses a challenge for businesses, as it becomes increasingly difficult to respond to them promptly and effectively.

One potential solution involves leveraging NLG models to automatically generate review responses. The recent advancements in NLG, particularly large pre-trained transformer-based language models, such as BART (Lewis et al. [2019]) or ChatGPT could offer a potential solution by enabling businesses to generate diverse, fluent, and timely responses to online reviews (Zhang et al. [2022]).

Despite these advancements, automating review response generation comes with its own set of challenges, the most prevalent one being the tendency for fine-tuned models to generate generic outputs. These are responses that are broad and universally applicable but lack specificity, which can lead to customer dissatisfaction and potentially harm a business's reputation. To overcome this, knowledge-enhanced generation offers a potential solution. This approach incorporates knowledge texts or attributes into the response generation process, intending to create outputs that are not only more specific but also informative, thus enhancing the experience for both the reviewer and prospective customers.

Yet, previous works have shown that merely incorporating additional information as input is not always sufficient to ensure the desired specificity. To address this, Hazarika et al. [2021] proposed a method for controllable knowledge-enhanced generation by applying an attention biasing knob in the attention mechanism of an encoder-decoder NLG model. The attention biasing knob adjusts the attention distribution during inference, aiming to improve the output quality such that it is more grounded and informative with respect to the appended knowledge. The effectiveness of their approach has been demonstrated in their study on a dialog generation domain, indicating that the attention biasing knob can enhance the informativeness of dialog responses.

Our motivation for this thesis lies in investigating the applicability of this attention biasing knob in knowledge-enhanced review response generation within the hotel domain. We frame the review response generation task in a similar manner to a dialog generation task, with the review serving as the first dialog turn and the response as the target turn. Given the shared challenge of genericness in both tasks, we believe that the attention biasing knob could be a viable approach to increase specificity and informativeness in review response generation.

## 1.2 Research Questions

The primary objective of this thesis is to investigate the applicability of the attention biasing knob, as presented in Hazarika et al. [2021], to review response generation within the hotel domain, with the aim of improving and controlling the informativeness of the generated output. Consequently, we formulate our first research question as follows:

**RQ1:** To what extent can cross-attention biasing be employed to generate more informative review responses given textual descriptions and details about a hotel?

In addressing this question, we explore the potential of applying attention biasing under two conditions:

- Experiment 1: A model fine-tuned on all available review response pairs with additional knowledge snippets, totaling approximately 2 million training examples.

- Experiment 2: A model fine-tuned on a ‘high-quality’ subset of review-response pairs, which is identified using the filtering approach described in Kew and Volk [2022].

This dual approach is motivated by findings from Kew and Volk [2022], suggesting that review response generation models tend to produce safe or generic responses, as they are prevalent in the training data and easy for the model to learn. Thus, we hypothesize that simply applying attention biasing during inference (Experiment 1) may not be sufficient to counter this tendency. Since our goal is to generate more specific responses that integrate additional details from the provided knowledge snippets, we anticipate that the second approach (Experiment 2) will yield a more adaptable model which can generate more specific responses using attention biasing.

For our first approach to RQ1, we take a simple approach that fine-tunes the BART model (Lewis et al. [2019]) using a dataset of approximately 2 million hotel review response pairs, incorporating two parts of knowledge as additional input (description and list of amenities of the hotel). During inference, we bias the model’s cross-attention weights towards these additional inputs.

The second approach of RQ1 fine-tunes the BART model again using a filtered training dataset, which corresponds to 40% of the entire dataset (approximately 800,000 instances), following the filtering methodology presented in Kew and Volk [2022]. This methodology filters out less generic and noisy data using perplexity scores. The attention biasing knob is applied as in the first approach.

Given the results of RQ1, we delve further into the potential of the attention biasing knob. Motivated by previous works that discussed the relationship between input length and model performances (Kew et al. [2020], Gao et al. [2021]), we pose an additional research question:

**RQ2:** How does the modification of knowledge input length during the inference stage impact the model’s performance with respect to informativeness?

We hypothesize that there exists an optimal knowledge input length range that triggers attention biasing as intended. Knowledge sequences that are too short may not contain sufficient enough to extract, even with appropriate bias values. Conversely, an excessively long knowledge input length may ‘*overwhelm*’ the model, causing it to struggle in focusing on the most pertinent information, thereby resulting in information dilution. In addressing RQ2, we conduct inference experiments utilizing the model trained for RQ1, while varying the knowledge input length during inference.

By answering our research questions, we aspire to contribute to the exploration

and advancement of more effective tools for automatic response generation, with potential implications in the hospitality industry for enhancing customer engagement, thereby improving customer satisfaction and the overall reputation of these businesses.

## 1.3 Thesis Structure

The structure of the thesis is as follows: In Chapter 2, we provide an extensive overview of the theoretical background and relevant prior works that underpin our research questions. Chapter 3 presents our preliminary experiment, which is designed to validate our implementation of the methodology by reproducing the results of the cross-attention biasing experiments from Hazarika et al. [2021]’s work. Addressing the first research question, Chapter 4 delves into a thorough presentation of the experimental procedures, setups, and resultant findings. In Chapter 5, we continue presenting the experiment associated with our second research question, along with its corresponding results. Chapter 6 offers an in-depth discussion of our findings, provides a comparative analysis with OpenAI’s ChatGPT, and acknowledges potential limitations while pointing to future research directions. Finally, Chapter 7 weaves together the various threads of our work, providing a comprehensive summary of our work and drawing final conclusions for this thesis.

## 2 Background

In this chapter, we provide an extensive overview of the theoretical background and previous works related to our research questions. Section 2.1 introduces tasks and recent development in Natural Language Generation (NLG). Section 2.2 elaborates on the review response generation task as a subfield of NLG. Section 2.3 expounds upon knowledge-enhanced generation, followed by recent works for a controllable knowledge-enhanced generation. Section 2.4 presents the Transformer architecture (Vaswani et al. [2017]) and approaches that aim to bias the attention module in NLG models. Section 2.5 introduces Hazarika et al. [2021]’s work on controlling the informativeness of the generated output using an attention biasing knob, which forms the basis of the experiments in this work. Lastly, Section 2.6 describes the automatic evaluation metrics for NLG tasks that we employed.

### 2.1 Natural Language Generation

Natural Language Generation (NLG) encompasses the generation of textual content in order to fulfill specified communicative goals (Dong et al. [2022]). Recent advancements in NLG techniques have facilitated the generation of fluent and more diverse sentences for a variety of tasks, such as summarization (Syed et al. [2021]), translation (Yang et al. [2020]), question answering (Chen et al. [2017]), dialog generation (Ghazvininejad et al. [2018]), and review response generation (Gao et al. [2019], Zhao et al. [2019], Gao et al. [2021]).

In the domain of NLG, sequence-to-sequence (seq2seq) encoder-decoder models, utilizing such as recurrent neural networks (RNNs), have been widely adopted. However, the advent of large pre-trained Transformer-based language models such as GPT-2 (Radford et al. [2019]), BART (Lewis et al. [2019]), or T5 (Raffel et al. [2020]) has ushered in a new era for NLG. These models have demonstrated considerable efficacy in generating coherent and contextually appropriate text. Owing to their pre-training on vast amounts of data, they exhibit remarkable versatility, allowing fine-tuning on smaller datasets for specific downstream tasks. More recent

larger pre-trained models such as GPT-3 (Brown et al. [2020]) and ChatGPT<sup>1</sup> are even capable of few-shot or zero-shot learning, indicating their aptitude to perform well on new generation tasks with minimal training data or even without any additional training. Despite these significant developments in NLG, generating outputs that conform to all requisite attributes without hallucination remains challenging. In the subsequent sections (Section 2.2 and Section 2.3), recent studies addressing these challenges will be discussed.

## 2.2 Review Response Generation

Online reviews offer customers a platform for exchanging their experiences and evaluations, which have been demonstrated to exert greater influence on product purchases than traditional marketing communications (Chevalier and Mayzlin [2006]). Engaging with online reviews not only enhances communication between customers and business entities but also provides tangible benefits for the latter. Industries such as mobile applications, e-commerce, and hospitality have demonstrated the advantages of providing responses to reviews. Hassan et al. [2018] reported that responding to an app review in the Google Play Store increases the likelihood of a user updating their given rating by up to sixfold in comparison to cases where no response is provided. In the e-commerce domain, evidence suggests that sellers who provide high-quality responses to reviews achieve a higher sales volume than their less responsive counterparts (Zhao et al. [2019]). The hospitality industry also benefits from engaging with reviews, as the success of firms in this sector relies not only on customer feedback but also on their ability to augment competitive advantages through customer interactions (Cantalops and Salvi [2014]). Consequently, hospitality businesses proactively respond to online reviews to foster customer engagement (Li et al. [2017]).

In light of the exponential growth of online reviews and their benefits for business entities, businesses are exploring strategies for efficiently engaging with reviews, leveraging NLG models as a means to automate response generation. One of the challenges in automated generation resides in circumventing generic, one-size-fits-all responses and producing more specific and informative output. This is particularly pertinent in the hospitality domain, where the informational value conveyed by long responses enables prospective customers to form better-informed evaluations of the establishment, ultimately leading to a more satisfactory experience (Li et al. [2017]). Example 2.1 depicts an instance of responses to a hotel review. **Response A** is

---

<sup>1</sup><https://chat.openai.com/>



rather generic and fails to address the nuanced information present in the review. This response could be applied to any review that conveys general satisfaction with the establishment. On the other hand, **Response B** is more specific and elaborates on the facility in greater detail, which is favored by customers. To improve specificity and informativeness, it is desirable to incorporate supplementary features that provide relevant contextual information for the response text instead of conditioning response generation solely on the review text (Kew et al. [2020]). Acknowledging the potential and challenges of automated review response generation, several studies have been undertaken with a particular emphasis on generating more specific and relevant responses.

### Example 2.1 — Generic and specific responses in a hotel domain

**Review:** “We stayed here for one night in January in a family cabin. Everything was clean, fresh and the park itself had a nice family atmosphere. Not far from the town center but nice and quiet. Recommend it as a stop-over or short-term.”

**Response A (generic):** “Thank you for taking the time to write about your stay with us, we’re really glad you enjoyed it. All the best and hope to see you again soon. Regards.”

**Response B (specific):** “Thank you for taking the time to write about your stay with us, we’re really glad you enjoyed it. The park is located two kilometers from the town center, which is about a minute in the car, or a leisurely 15-minute walk along the walking track around the bay. so, like you mentioned: close to town but just far enough away to be peaceful and quiet. All the best and hope to see you again soon. Regards.”

Gao et al. [2019] proposed RRGGen, a model that extends an encoder-decoder model with an attention mechanism for the automated generation of responses to app reviews. The model incorporates review attributes such as user rating, review length, sentiment score, and app category to capture user sentiment and topics in order to generate more relevant responses. Their qualitative analysis demonstrated that RRGGen effectively generates relevant and accurate responses, outperforming baseline methods. Kew et al. [2020] adapted this approach for the hospitality domain, incorporating features such as sentiment score, review categories, and a keyword dictionary, akin to Gao et al. [2019]. However, they discovered that the adapted approach encountered difficulties in generating suitable review responses when applied to the hospitality domain. They suspected that the discrepancies in source and target text lengths between the app dataset in Gao et al. [2019] and the hospitality dataset might be the underlying cause. Additionally, their hospitality dataset exhibited a higher degree of uniqueness (94%) compared to the app dataset (40%).

Zhao et al. [2019] presented another approach to incorporate additional attributes into the generated response by adapting a seq2seq model (Sutskever et al. [2014]) for e-commerce review response generation. Their architecture leverages product meta-information such as brand, material, color, and size, with the aim of generating product-specific, high-quality responses. A gated multi-source attention mechanism and a copy mechanism were utilized to leverage the product information.

Katsiuba et al. [2022] fine-tuned BART for review response generation in the hospitality domain. They prepend discrete tokens during the fine-tuning and post-process the generated responses by replacing the discrete tokens with suitable phrases and named entities for a given review and response pair, such as business-specific signature greetings and salutations. This approach aimed to generate more specific and personalized responses. Although the generated responses were grammatically correct and internally coherent, they tended to produce generic responses and hallucination was reported as well.

## 2.3 Knowledge-enhanced Text Generation

Incorporating knowledge sources into generation systems is referred to as knowledge-enhanced text generation, with the objective of generating more specific desired output. Yu et al. [2022] divided different knowledge sources into two categories (see Figure 1): internal knowledge and external knowledge. Internal knowledge is derived from within the input texts, such as keywords, topics, and linguistics features. Gao et al. [2019] and Kew et al. [2020], as discussed in Section 2.2, leverage internal knowledge sources extracted from input texts, such as review length, keyword, or sentiment scores for a knowledge-enhanced generation.

In contrast, external knowledge is obtained from outside sources, such as a knowledge base, knowledge graph, or grounded text. Incorporating external knowledge can provide a broader range of information and can lead to more diverse and specific output compared to incorporating internal knowledge sources. The grounded text refers to textual information that can provide additional knowledge relevant to input sequences (Yu et al. [2022]). It can be obtained from online resources such as encyclopedias, social media, or domain-specific web sources like description texts of services or products (Amazon reviews).

A substantial body of work has leveraged grounded text for response generation. Although not within the domain of review response generation, Ghazvininejad et al. [2018] adapted a seq2seq model for dialog generation by conditioning responses on

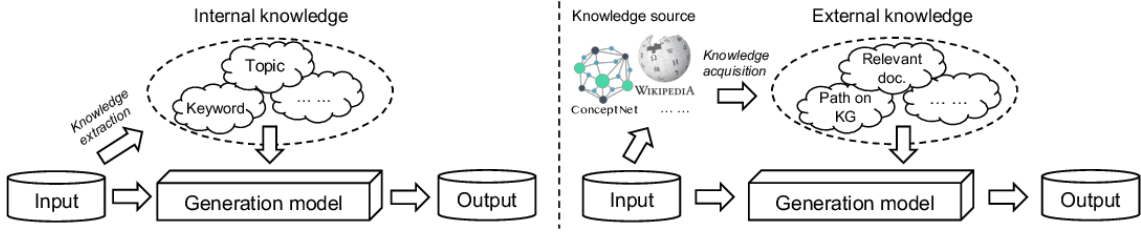


Figure 1: Two types of knowledge sources for knowledge-enhanced text generation. Figure from Yu et al. [2022]

both conversation history and external grounding text to produce more contentful responses. They utilize Foursquare, Wikipedia, or Amazon Reviews as knowledge sources, extracting contextually relevant facts (serving as external grounding text) via an entity name matching method. These facts are then fed as additional input along with conversation history.

Similarly, Dinan et al. [2018] employed the Memory Network (Sukhbaatar et al. [2015]) and Transformer for grounded dialog generation, using Wikipedia as a knowledge source. Through Information Retrieval (IR) and attention mechanisms, they retrieve and select which knowledge sentence from Wikipedia should be used as additional input for grounding. De Bruyn et al. [2020] expanded upon the work of Dinan et al. [2018], using the same dataset. They fine-tuned a BART model, enabling it to retrieve multiple sentences, in contrast to Dinan et al. [2018], which limits knowledge input to a single sentence. By feeding the model multiple sentences, it was designed to further select the relevant information part in the decoder.

Gao et al. [2021] further developed their previous work (Gao et al. [2019]) for review response generation by incorporating other knowledge sources. The new architecture, called Contextual knowledge-based app Review response generation (CoRe), addresses the potential negative impact of using keywords or sentiments as additional knowledge on the architecture’s results because the model performance depends on the accuracy of keywords/sentiments extraction tools. To mitigate this limitation, Gao et al. [2021] incorporates external app descriptions instead of using information directly extracted from the input. Furthermore, they noted that the previous system favored high-frequency tokens in the corpus, resulting in the generation of responses that were often generic and uninformative. As a proposed solution, the new system incorporates responses of similar reviews, utilizing the IR approach (Ji et al. [2014]) to generate low-frequency tokens as well. CoRe surpassed its previous architecture, RRGGen, in all evaluation metrics.

### 2.3.1 Controllable Knowledge-enhanced Generation

The majority of the literature discussed thus far has primarily focused on methods for retrieving pertinent knowledge to serve as additional input for grounding the response. For instance, Dinan et al. [2018] and Gao et al. [2021] employed information retrieval techniques to select the most suitable knowledge snippets. However, the black-box nature of complex neural networks, including pre-trained language models, makes controlling the output challenging. Consequently, more recent studies have shifted their focus towards devising strategies for controlling knowledge-enhanced text generation. Although conditioning generation on grounded text can enhance information reliability and inhibit the hallucination of facts to some extent, it may also result in vague or irrelevant responses. Introducing controllability on top of this can yield contentful and reliable responses (Wu et al. [2021]).

Wu et al. [2021] proposed a framework for controllable grounded response generation in dialog systems. During the fine-tuning of GPT-2-based models, control phrases are appended as supplementary input to guide the generated text towards incorporating semantically relevant information from the knowledge source that corresponds to the control phrase. The approach can be characterized as a form of soft semantic control, as it aims to extract specific semantic aspects from the knowledge rather than merely allowing the model to select relevant information.

Rashkin et al. [2021] similarly explored controllable grounded dialog response generation. They append various control codes as additional tokens during the fine-tuning of transformer-based models. These control codes represent the extent to which the response is grounded in the knowledge part. During the decoding process, a resampling method is employed to generate outputs that adhered to the desired control code. The primary objective is to control groundedness, also referred to as faithfulness in their study, in the generated response.

One limitation of both of these approaches is that the model needs to be trained with control codes or tokens, which makes them less flexible and computationally expensive. In contrast to those approaches, Hazarika et al. [2021]’s approach for knowledge-enhanced dialog generation enables soft controlling in a zero-shot manner that does not require any control-specific training. A detailed introduction to this approach will be provided in Section 2.5.

## 2.4 Transformer

A significant portion of large pre-trained language models is based on the Transformer architecture (Vaswani et al. [2017]), a revolutionary neural network architecture developed to address the limitations of traditional sequence-to-sequence models, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), which process input sequences in a sequential manner. The transformer relies exclusively on attention mechanisms, eliminating the recurrence and convolutions entirely. It consists of an encoder and a decoder, and its attention mechanism enables parallel processing of input sequences.

BART(Lewis et al. [2019]), which is employed in our experiment, is one of the large pre-trained language models constructed on a standard Transformer architecture. It is trained by corrupting documents and then optimizing a reconstruction loss, which is the loss between the decoder’s output and the original, uncorrupted document. The autoregressive decoder of the BART enables direct fine-tuning of the model for sequence generation tasks.

OpenAI’s<sup>2</sup> ChatGPT, another large language model, is a variant of the Generative Pre-trained Transformer (GPT) series. ChatGPT is fine-tuned specifically for conversational applications leveraging Reinforcement Learning from Human Feedback (RLHF), making it suitable for generating conversational responses.

In Section 6.2, we present a comparative analysis of our BART-based method investigated in this thesis against ChatGPT, aiming to provide a comprehensive understanding of our proposed approach and to gain deeper insights into these Transformer-based models.

### 2.4.1 Transformer Architecture

The Transformer architecture, depicted in Figure 2, comprises an encoder and a decoder, each consisting of stacked layers containing feed-forward networks and multi-head attention. The encoder processes an input sequence and generates a continuous representation, while the decoder receives the encoder’s output and generates the output sequence in an autoregressive manner. Positional embeddings preserve the order of tokens in the sequence. Layer normalization and residual connections are employed to ensure stable training.

---

<sup>2</sup><https://openai.com/>

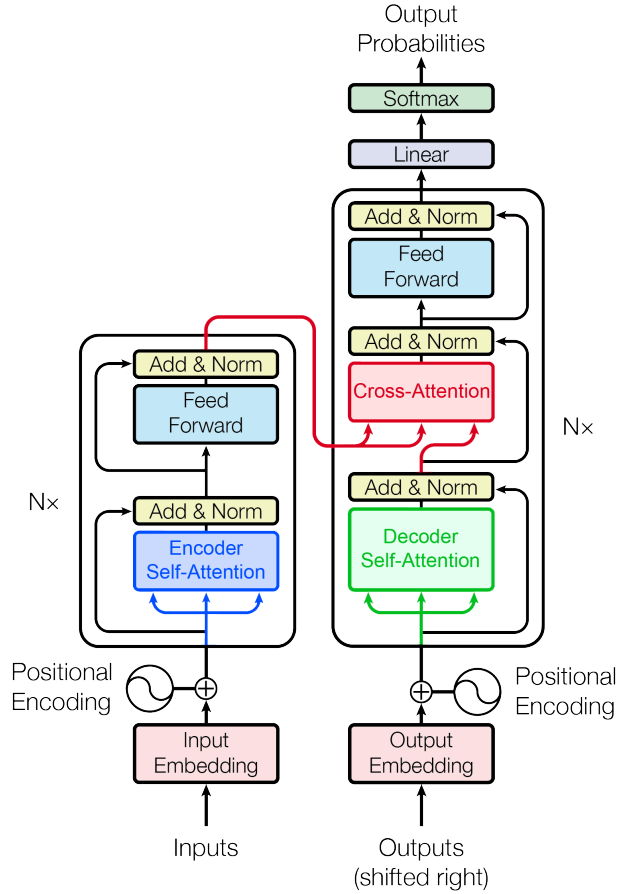


Figure 2: The Transformer architecture. Figure adapted from Vaswani et al. [2017].

The attention mechanism is a vital component of the Transformer. The Transformer incorporates three types of attention: encoder self-attention, decoder self-attention, and cross-attention. As illustrated in Equation 2.1, an attention function maps query( $Q$ ), key( $K$ ), and value( $V$ ) vectors, which are derived from the output of previous layers in the model, to an output. In the scaled dot-product attention used in Transformer, the attention score  $QK^T$  is determined by the dot product of the query matrix  $Q$  and the key matrix  $K$ , with the score scaled down by  $\sqrt{d_k}$ . Subsequently, the softmax function is applied to the scaled attention score to produce the attention distribution ( $\text{softmax}(\frac{QK^T}{\sqrt{d_k}})$ ). Finally, the attention distribution is multiplied by the value matrix to yield the attention value. This attention value is employed to determine the amount of attention to be allocated to each context token.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

**Encoder self-attention** (highlighted in blue in Figure 2) enables each token in the encoder to attend to all the other tokens in the encoder. In other words, it allows the model to evaluate the significance of different tokens in the input sequence relative to each other. This mechanism assists in capturing long-range dependencies and rich bidirectional context in the input sequence.

**Decoder self-attention** (highlighted in green in Figure 2) permits each token in the decoder to attend to all previously generated tokens in the decoder stack. To ensure that each token attends only to the prior tokens in the stack, a look-ahead mask is employed. The decoder’s attention allows the model to consider the context of the previously generated tokens while generating the next token in the output sequence.

**Cross-Attention** (highlighted in red in Figure 2), also known as encoder-decoder attention, enables the decoder to attend to various parts of the input sequences in the encoder at every generation timestep. This mechanism is similar to the attention mechanisms employed in traditional sequence-to-sequence models. Cross-attention allows the decoder to incorporate relevant information from the input sequence while generating the output sequence.

## 2.4.2 Biasing Attention Module

Recognizing the importance of the attention mechanism in the Transformer architecture, several studies have been conducted to modify and bias the attention, primarily aimed at achieving focused attention, particularly for neural machine translation tasks.

One method, proposed by Shaw et al. [2018], extends the self-attention mechanism to incorporate representations of the relative positions of tokens in the sequence, as opposed to using absolute positions as used in the original Transformer. Another method, introduced by Yang et al. [2018], employed a learned Gaussian bias term in the attention score calculation to improve the Transformer model’s ability to capture local information in self-attention networks. This term biases the attention weights towards neighboring tokens, increasing the model’s sensitivity to the local context. They also conducted experiments with a combination of the learned Gaussian bias and relative position representations (Shaw et al. [2018]), reporting that the combination yielded better results than applying a single variant. You et al. [2020] also introduced a Gaussian bias to the attention calculation. However, unlike in Yang et al. [2018], this method utilizes hard-coded Gaussian bias without any learned parameters, and all attention heads are replaced with hard-coded attention. They

demonstrated that the hard-coded attention variant in decoding and encoding self-attention has minimal impact on BLEU scores in machine translation tasks, whereas completely replacing cross-attention with the hard-coded variant significantly lowered the BLEU scores.

## 2.5 Cross-Attention Biasing in Knowledge-enhanced Generation

This section presents an overview of the methodology introduced by Hazarika et al. [2021], which serves as the primary approach for this work. The study proposes a technique aimed at improving and controlling the informativeness of generated outputs in knowledge-enhanced generation tasks. They achieve this through the use of cross-attention biasing, referred to as the ‘attention biasing knob’ in their paper. This approach involves directly modifying the attention distribution to adjust the allocation of attention to different parts of the input, through the adjustment of a single hyperparameter. The underlying assumption is that increasing attention to the grounding knowledge snippet would lead to more informative and specific outputs, as the output incorporates the knowledge to a higher degree. Example 2.2 demonstrates how grounding a response to the provided knowledge snippet could make it more informative, as information from the snippet is extracted and incorporated into the response.

### Example 2.2 — An example of knowledge-enhanced response generation, taken from Hazarika et al. [2021]

**Previous turns:**

**A** “Hi! do you like to dance?”

**B** “I love to dance a lot. How about you?”

**A** “I am really bad, but it is a good time.”

**Knowledge:** Bruce Lee was also a great dancer and that he won the Hong Kong Cha-Cha Championship in 1958.

**Response (next turn) :**

**Uninformative** **B** “Hmm. Dancing is a lot of fun.”

**Informative** **B** “Dancing is a lot of fun. Even Bruce Lee was a great dancer and has won competitions.”



Figure 3 provides a visual representation of the cross-attention biasing knob. Hazarika et al. [2021] introduces a bias vector  $\mathbf{b}_t$ , composed of bias values for each input token. In this figure, a bias value of 5 is assigned to the green part of the input in order to generate output more relevant to the part. The attention distribution is adjusted according to the bias vector, resulting in a generally higher value for the green part of the input. All the biasing calculations are applied equally across multiple heads of attention layers.

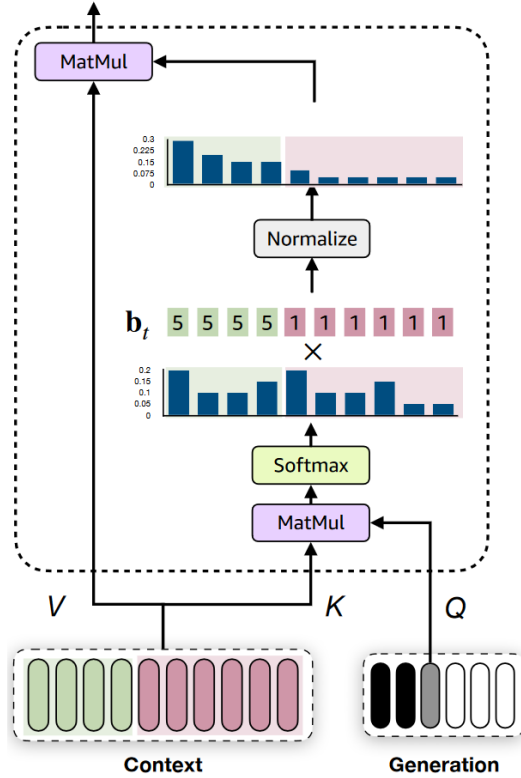


Figure 3: Visualization of the attention biasing knob for cross-attention. Bias values of 1 (pink) and 5 (green) are allocated to the different parts of the input. Figure from Hazarika et al. [2021]

Equation 2.2 depicts the modified attention calculation for the attention biasing knob. Given attention matrices  $K$ ,  $V$ ,  $Q$ , and bias vector  $\mathbf{b}_t$ , the biased dot product attention is computed by element-wise multiplication of the bias vector  $\mathbf{b}_t$  with the attention distribution. The modified distribution is then re-normalized to produce a valid probability distribution. This new distribution is referred to as ‘biased attention distribution’ (refer to Equation 2.1 for the original attention distribution).

$$Attention(Q, K, V, \mathbf{b}_t) = \mathcal{N}(\mathbf{b}_t \odot softmax(\frac{QK^T}{\sqrt{d_k}}))V \quad (2.2)$$

Hazarika et al. [2021] conducted experiments with the attention biasing knob on dialog generation tasks and demonstrated an increase in output informativeness. Again, one advantage of this approach is its ability to control models during inference in a zero-shot manner, without the need for additional control-specific training. Unlike other controlling approaches introduced in Section 2.3.1, this method does not require gradient updates for output control. Another advantage is the applicability of this method to any encoder-decoder transformer model, allowing for easy extension to other applications. However, it is essential to note that this approach is intended to be applied to fine-tuned models, where the knowledge snippet is appended as additional input.

We reproduce the findings from Hazarika et al. [2021] as preliminary experiments on a dialog generation task (Section 3) and conduct our main experiments on a review response generation task (Section 4 and 5).

## 2.6 Evaluation Metrics for Natural Language Generation

The evaluation of an NLG system is typically conducted using automatic evaluation metrics and/or human evaluation. While human evaluation allows for more task-specific and fine-grained assessments, it is often expensive, and quality control is crucial. In contrast, automatic evaluation metrics are more generic but easier to implement and reproduce. Depending on the tasks and objectives of the NLG system, the selection of appropriate metrics is critical. Inadequate metrics can lead to invalid evaluations of the system. Consequently, evaluation metrics for NLG have long been a research focus, with researchers developing task-specific metrics or evaluating existing ones. In this section, we introduce the automatic evaluation metrics employed in our experiments.

### 2.6.1 Reference-based Metrics

Reference-based metrics evaluate machine-generated texts with respect to reference texts, which are typically human-generated. The Bilingual Evaluation Understudy (**BLEU**) metric (Papineni et al. [2002]) calculates the n-gram matches between the generated and reference texts. Originally designed for evaluating machine translation models, BLEU has since been applied to other NLG tasks, including dialog generation. This metric emphasizes precision, so shorter predicted sentences attain

higher scores than longer ones. To counteract this, a brevity penalty is introduced, which penalizes predictions that are too short compared to the reference length.

The Recall-Oriented Understudy for Gisting Evaluation (**ROUGE**) metric (Lin [2004]) is another n-gram-based metric that focuses on recall. ROUGE-N measures N-gram co-occurrences, while ROUGE-L quantifies the longest common subsequence between two texts. It is commonly employed in text summarization and other NLG tasks.

The Metric for Evaluation of Translation with Explicit Ordering (**METEOR**) metric (Banerjee and Lavie [2005]) was developed to address the weaknesses of the BLEU metrics, such as its lack of recall and explicit word-matching (Dong et al. [2022]). METEOR matches unigrams based on surface forms, stemmed forms, and meanings, and calculates the score based on the harmonic mean of precision and recall, with a slight emphasis on recall.

While BLEU, ROUGE, and METEOR compute n-grams at the word level, **ChrF** (Popović [2015]), calculates n-gram scores at the character level, balancing precision and recall. **ChrF++**(Popović [2017]) additionally incorporates word n-grams for scoring, resulting in better correlation with human judgments. ChrF is considered to be more flexible than BLEU, particularly for noisy web-based text data with spelling errors (Kew and Volk [2022]).

**Unigram precision and recall** can be used as a simple and effective metric to calculate the lexical overlap between the generated and reference texts. Precision is the proportion of unigram overlap out of all unigrams in the generated output, measuring the correctness of the generated text. Recall, on the other hand, assesses how many unigrams in the reference text are captured in the generated text, calculated by the proportion of unigram overlap out of all unigrams in the reference text. The F1 score, the harmonic mean of these two, then harmonizes these two aspects to provide a comprehensive metric for evaluation.

The metrics introduced above are untrained, lexical-based evaluation metrics, easily implementable without additional training cost. In contrast, **BERTScore** (Zhang et al. [2019]) is a semantic-based machine-learned evaluation metric designed to closely simulate human judgment (Dong et al. [2022]). It utilizes pre-trained contextual embeddings from BERT (Devlin et al. [2018]) to measure the semantic similarity between two texts, allowing for the calculation of recall, precision, and F1 scores.

## 2.6.2 Reference-free Metrics

**Perplexity** (PPL) is a metric employed to measure the performance of a language model, reflecting its ability to predict a given sample. It is often utilized in NLG tasks as a proxy for assessing the fluency of texts using a pre-trained language model. In this context, lower perplexity values indicate better fluency. However, perplexity is often biased toward shorter sentences, and less frequent words tend to exhibit worse perplexity than frequent words with the same meaning (Jin et al. [2022]). It is also important to note that the effectiveness of perplexity as a fluency measure is not without contention. Some studies, such as one by Mir et al. [2019], suggest that there is no significant correlation between perplexity and human fluency ratings, bringing into question the universality of its application.

**Self-BLEU** (Zhu et al. [2018]) gauges the diversity of generated texts. It calculates the BLEU score of each generated text against other texts, then defines the average score as the Self-BLEU score. A lower Self-BLEU score signifies higher diversity.

## 2.6.3 Evaluation Metrics for Knowledge-enhanced Generation

The reference-based metrics introduced in Section 2.6.1 typically use human-generated gold target texts as references. In the case of the review response generation task, the reference texts are human-written responses paired with the input reviews. However, depending on the objective of the task, it is possible to set the reference text differently. In Kew and Volk [2022], which aimed to enhance specificity in review response generation, they computed ChrF twice, once using gold references as standard and once using the corresponding reviews as references to provide an approximate measure of specificity in model outputs.

In the context of knowledge-enhanced generation, where the evaluation aims to assess groundedness or informativeness, evaluating the generated texts based on the human response may not provide insights into these aspects. Therefore, some studies have utilized reference-based evaluation metrics with respect to grounding knowledge. For instance, Rashkin et al. [2021] measured the precision and recall of the unigrams in the response with knowledge texts as a reference, expecting that a high value would indicate that a high proportion of words in the generated response are contained in the knowledge text. Similarly, Hazarika et al. [2021] calculated BLEU, ROUGE, and METEOR of the generated texts with respect to the relevant knowledge snippet to evaluate the informativeness of responses. We adopt the same approach to measure informativeness in our review response generation task.

## 3 Preliminary Experiment

In this chapter, we describe a preliminary experiment aimed at reproducing the results of the cross-attention biasing experiments from Hazarika et al. [2021]. This experiment entails modifying the cross-attention distribution during inference to adjust the attention allocated to different parts of the input. The motivation for conducting this preliminary experiment is to ensure the correctness and validity of our implementation of the methodology from the original paper before moving on to our primary experiments. Section 3.1 delves into the dataset and experimental setup<sup>3</sup>, closely following the methodology described in Hazarika et al. [2021]. In Section 3.2, we present the results and draw comparisons with the scores obtained in the original study. Section 3.3 offers a conclusion for the preliminary experiment.

### 3.1 Experimental Setup

#### 3.1.1 Dataset

Hazarika et al. [2021] employed the Topical Chat dataset, originally introduced in Gopalakrishnan et al. [2019]. This dataset includes dialog between two Mechanical Turk workers (also known as Turkers) discussing specific topics. During the dialogs, Turkers were provided with a ‘reading set’ containing knowledge extracted from various data sources, such as Wikipedia, Reddit, and the Washington Post, which served as reference material for knowledge-enhanced conversations. Prior work by Hedayatnia et al. [2020] has aligned the appropriate knowledge snippet in the reading set and the ground truth target response using semantic similarity. Using these alignments, the knowledge snippets with the highest similarity were selected as additional input for a knowledge-enhanced generation.

---

<sup>3</sup>We used Tannon Kew’s attention biasing implementation in this repository: <https://github.com/tannonk/transformers>. Other code implementations are further developed based on the source code in this repository: <https://github.com/ZurichNLP/understanding-ctx-aug>

### 3.1.2 Formatting the Input

We adhered to the input construction method described in the original paper. Example 3.1 presents a specific instance of a single input from the Topical Chat dataset utilized in our experiment. One input consists of five dialogs **Turns** and an additional **Knowledge** snippet. The **Target** is the subsequent turn (the 6th turn) of the input turns. The underlined parts in the knowledge and target in the example represent the similarity between the two sections.

#### Example 3.1 — A single input of the Topical Chat dataset

**Turns:**

**A** “Did you know that the University of Iowa’s locker room is painted pink? I wonder why?”

**B** “I think I did hear something about that. I imagine it is an attempt to psych the other team out.”

**A** “So, it would be in the visiting team’s locker room but not their own?”

**B** “Right. Teams do all kinds of things to bother the competition. I’ve heard of teams having heated benches in the winter for themselves but not for the visitors.”

**A** “I would hate a cold bench. Then again, I wouldn’t want to be some place that cold or watching football.”

**Knowledge:** That the highest score ever in a football game occurred in 1916 when Georgia Tech defeated Cumberland 222-0

**Target (ground truth response) :**

**B** “I’d rather watch it inside where it’s warm. Have you heard about the Georgia Tech-Cumberland game of 1916?”

In accordance with Hazarika et al. [2021], we fixed the input length by assigning a fixed number of tokens for each bucket. A bucket, in this context, refers to each input section, such as a single dialog turn and a knowledge snippet. Figure 4 displays the token length percentiles of the turn, knowledge, and target bucket in the entire dataset. The bucket sizes defined by the original study approximately correspond to the 75th percentile of the dataset. Concretely, the bucket size of the knowledge snippet was fixed to 32 tokens, and each turn bucket was assigned 30 tokens. If the number of tokens in a bucket exceeded the bucket size, the text in the bucket was truncated. Conversely, if the number of tokens in a bucket was less than the bucket size, the bucket was padded to the bucket size on the right.

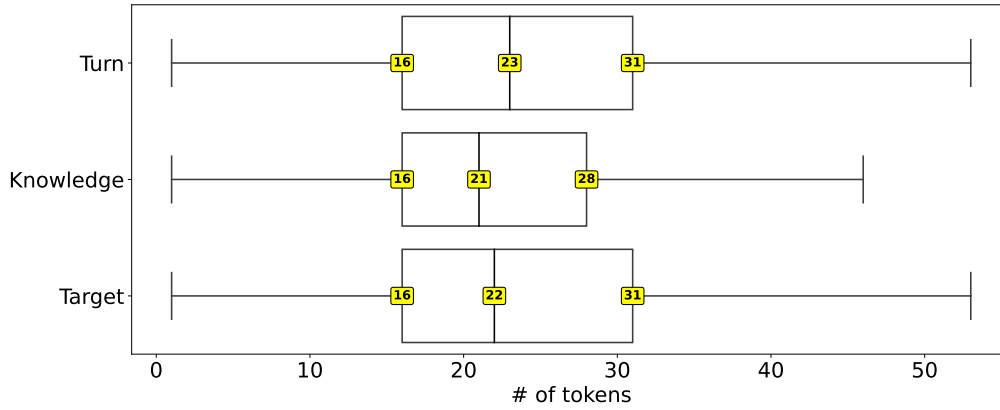


Figure 4: The 25th, 50th, and 75th percentiles of the turn, knowledge and target token lengths in the Topical Chat dataset (including the train, validation, and test sets) are presented. Note that a turn here represents a single dialog turn. The numbers in yellow correspond to the token lengths at the 25th, 50th, and 75th quantiles, respectively. Outliers have been excluded from this analysis.

Special tokens were added as illustrated in Example 3.2. Hazarika et al. [2021] appended the beginning of a sentence (BOS) token `<s>` and speaker identification tokens `<speaker1>`, `<speaker2>` for each turn to mark boundaries between turns. In our experiment, we added the end of a sentence (EOS) token `</s>` along with the input category identifiers (`<knowledge>` and `<turns>`) to the input format.

### Example 3.2 — Input Formatting of Topical Chat Dataset

**Input formatting in the Hazarika et al. [2021]:**

```
<s> knowledge <speaker1> turn1 <speaker2> turn2 ...
```

**Input formatting for our preliminary experiment:**

```
<s> <knowledge> knowledge <turns> <speaker1> turn1 <speaker2> turn2 ... </s>
```

Consequently, each input length was fixed at 191 (knowledge snippet (32) + five turns (30\*5) + special tokens(9)). Note that the total input length slightly differs from that in Hazarika et al. [2021] because we explicitly include separate special tokens to represent knowledge and turns. This allows for better generalization to other datasets and tasks. Furthermore, unlike in the original study, we do not add our special tokens to the vocabulary. As a consequence, these tokens are split into multiple tokens during tokenization, e.g. `<speaker1>` into `{<, speak, er, 1, >}`.

The number of data points used in our experimentation is presented in Table 1. The validation and test datasets consist of two sets: frequent set (**valid\_freq**, **test\_freq**), and rare set (**valid\_rare**, **test\_rare**). The frequent set encompasses entities frequently observed in the training set, while the rare set consists of entities scarcely seen in the training set (Gopalakrishnan et al. [2019]). For our preliminary experiment, we only employed the frequent sets as our validation and test data. However, Hazarika et al. [2021] showed that the results of frequent and rare sets are comparable.

		# of data	# of unique knowledge	Uniqueness ratio
	Train	145,238	23,655	0.16
Validation	valid_freq	8,986	1,643	0.18
	valid_rare	8,997	2,674	0.30
Test	test_freq	9,065	1,220	0.13
	test_rare	9,075	2,840	0.31

Table 1: Overview of the Topical Chat dataset. The frequent set comprises entities that are commonly observed in the training data, while the rare set encompasses entities that appear infrequently in the training data. The uniqueness ratio is calculated by dividing the number of unique knowledge snippets by the number of datapoints in the dataset.

### 3.1.3 Fine-tuning and Inference

Consistent with the experimental configuration in Hazarika et al. [2021], we fine-tuned the  $\text{BART}_{\text{BASE}}$  model on our dataset for ten epochs, employing a learning rate of  $6.25\text{e-}5$ . For inference, we utilized nucleus sampling (Holtzman et al. [2019]) with a top-p value of 0.9, and the temperature setting of 0.7. The maximum response (target) length was configured to 40 tokens. Furthermore, bias profiles for adjusting the level of control were provided as additional inference arguments. We conducted experiments with two distinct biasing profiles (**Knowledge**, **Dialog**), as depicted in Figure 5. The profiles were configured as follows:

**Dialog:** In the dialog biasing profile, the knowledge bias was fixed at 1, while a higher bias value was assigned to the dialog part. Specifically, we set the bias values to  $(b_t^k, b_t^h) = (1, 5)$  for all timesteps.

**Knowledge:** In the knowledge biasing profile, the dialog bias was fixed at 1, and an increased value was given to the knowledge bias. We experimented with various knowledge bias values to examine the level of control as the bias



value escalated. Specifically, we set  $(b_t^k, b_t^h) = (\delta, 1)$  where  $\delta \in \{2, 5, 10, 50\}$  for all timesteps.

**No Bias (Baseline):** The baseline without attention biasing was set to  $(b_t^k, b_t^h) = (1, 1)$  for all timesteps.

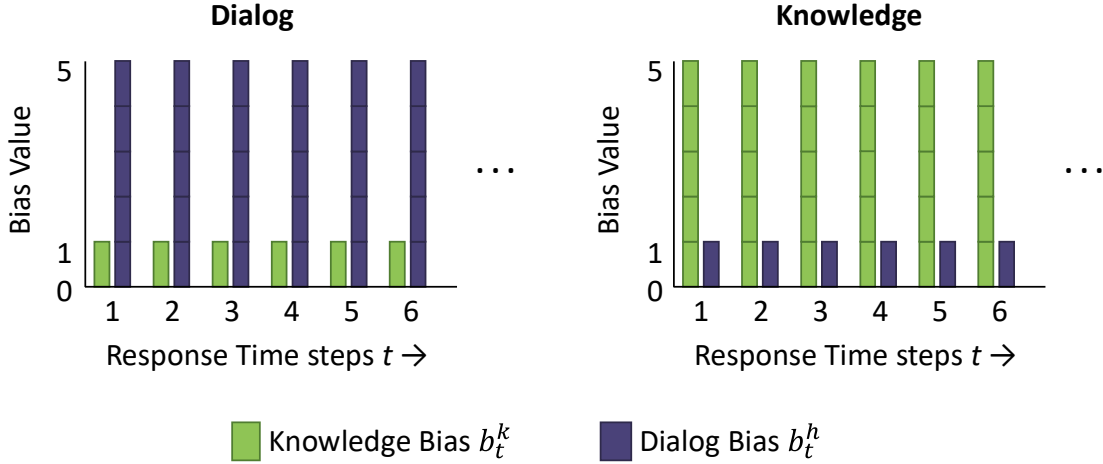


Figure 5: Two cross attention biasing profiles. In the dialog profile (left), the bias value for the dialog part (turns) is set to 5 for every timestep, while the bias value for the knowledge part is assigned a value of 1. In contrast, the knowledge profile on the right side biases knowledge part, allocating the value of 5 for every timestep. Figure adapted from Hazarika et al. [2021].

## 3.2 Results

### 3.2.1 Evaluating Informativeness

In this task, we interpret ‘informativeness’ as the extent to which information from the provided knowledge snippet is integrated into the generated responses. The degree of informativeness can be then quantified by gauging the n-gram overlap between the generated responses and the knowledge snippet. Hazarika et al. [2021] employed reference-based metrics such as BLEU, ROUGE, and METEOR, which incorporate n-gram overlap calculations. As the goal is to measure informativeness, these scores are calculated with respect to knowledge snippets, not ground truth. These metrics were also adopted in our preliminary experiment to evaluate our output performance in terms of informativeness. Nonetheless, as automatic metrics are not always reliable in assessing the informativeness of the response (Belz and Reiter [2006]), a human evaluation was conducted additionally in Hazarika et al. [2021].

Due to resource constraints, we excluded human evaluation in our reproduction experiment.

Table 2 presents a summary of the automatic evaluation metrics from our implementation and Hazarika et al. [2021]<sup>4</sup>. Again, it is important to note that the reference texts for these metrics are knowledge snippets ( $k$ ) to evaluate the informativeness of the generated text. The results reported in Hazarika et al. [2021] (marked in grey in the table) reveal that applying the cross-attention bias to the dialog (**Dialog 5**) led to lower BLEU $_k$ , ROUGE $_k$ , and METEOR $_k$  scores, indicating that the provided knowledge snippet is incorporated less in the generated responses. Conversely, when the bias was applied to the knowledge part (**Knowledge N**), these metrics were higher in comparison to the baseline. In particular, ROUGE $_k$  showed that a higher knowledge bias resulted in increased n-gram recall from the knowledge snippet. Our reproduction experiment exhibits the same trend. When the bias was applied to the dialog, the scores dropped compared to the baseline. Conversely, when the biases are applied to the knowledge part, all metrics showed improvement. Overall, the stronger the bias on the knowledge snippet, the higher the n-gram overlap scores.

<b>Bias Profile</b>	<b>BLEU<math>_k</math></b>	<b>BLEU<math>_k</math></b>	<b>ROUGE-1<math>_k</math></b>	<b>ROUGE<math>_k</math></b>	<b>METEOR<math>_k</math></b>	<b>METEOR<math>_k</math></b>
No bias (Baseline)	0.13	0.09	0.29	0.22	0.28	0.28
Dialog 5	0.06	0.03	0.19	0.13	0.18	0.16
Knowledge 2	0.17	-	0.35	0.25	0.33	-
Knowledge 5	0.22	0.14	0.42	0.28	0.40	0.36
Knowledge 10	0.25	-	0.46	0.32	0.45	-
Knowledge 50	0.32	-	0.55	0.38	0.53	-

Table 2: Automatic evaluation metrics for the preliminary experiment conducted on the Topical Chat dataset (frequent test set). The scores are averaged across five seeds. For comparison, the scores on the right side are the results from Hazarika et al. [2021] (corresponding to Table 3 and Table 6 in the original paper). The scores do not align precisely due to discrepancies in the experimental setup, which can include variations in random seeds and slight differences in input formatting. Note that Hazarika et al. [2021] did not specify which n-gram was used for the ROUGE score, and their scores are only partially available.

<sup>4</sup>Note that in their original study, Hazarika et al. [2021] conducted an experiment with a bias value of 5 and reported all metrics. In a subsequent experiment, which aimed to measure the effect of varying bias amounts, they only employed ROUGE as their evaluation metric. Thus, their scores are only partially available in the table.

### 3.2.2 Evaluating Fluency

Fluency refers to the grammatical and syntactical correctness of the generated responses. Hazarika et al. [2021] investigated whether applying attention biasing to produce more informative responses would adversely affect fluency. To measure fluency, they included human evaluations along with perplexity.

Table 3 presents the perplexity scores of the generated texts. In Hazarika et al. [2021], perplexity exhibited a slight increase as the model was biased with different profiles. However, Hazarika et al. [2021] reported that human evaluation on fluency showed no significant difference between the baseline and the models with biasing profiles, indicating that applying attention biasing does not significantly impair fluency. Our experiment findings demonstrate a similar pattern. When the bias was applied to the knowledge, perplexity slightly increased. However, when the bias was given to dialog with the value of 5, perplexity decreased relative to the baseline, which is a different behavior from Hazarika et al. [2021]’s result. While the original study mentioned that perplexity was computed with respect to human responses ( $PPL_r$ ), it did not offer further specifics about the calculation process. In our case, the perplexity scores of each generated response were calculated using the DistilGPT2 model (Sanh et al. [2019]). The slight deviation in the results could be attributed to the methodological discrepancy.

Bias Profile	PPL	$PPL_r$
No bias	85.76	9.66
Dialog 5	73.48	10.15
Knowledge 2	93.33	9.78
Knowledge 5	103.37	10.20
Knowledge 10	109.73	10.70
Knowledge 50	117.58	12.23
Ground truth	125.07	-

Table 3: Perplexity scores for the preliminary experiment conducted on the Topical Chat dataset (frequent test set). The scores are averaged across five seeds. The scores on the right side are the results from Hazarika et al. [2021]. Note that a direct comparison of the scores is not feasible, as perplexity computation differs between the two experimental setups.

### 3.3 Conclusion

In an effort to validate and expand upon the findings of the original study by Hazarika et al. [2021], who aimed to investigate the controllability of the informativeness of the generated output by allocating higher attention to knowledge-enhanced text, we conducted a reproduction experiment following the same methods and procedures as closely as possible. Observations from our experiment exhibit trends that are similar to those reported in the original study. Applying bias to dialog history lowered the informativeness scores compared to the baseline while applying it to knowledge snippets improved all metrics. Fluency followed a comparable pattern, with perplexity slightly increasing when biasing towards knowledge. We observed a minor discrepancy in perplexity when biasing towards dialog, which may be attributable to model differences in perplexity calculation. Despite this discrepancy, the overall pattern of the results aligns with the original study, indicating that the conclusions drawn from both experiments are largely in agreement. By replicating the results from the original study, we demonstrated that our implementation is accurate and reliable, thus providing a solid foundation for our subsequent experiments.

## 4 RQ1 Experiments

In this chapter, we present our experiments designed to address our first research question. In Section 4.1, we introduce the attention biasing methodology employed across all our experiments, along with the data filtering methodology specific to our second experiment under RQ1. Section 4.2 delves into the details of our experimental setup, outlining the dataset and the procedures for fine-tuning, inference, and evaluation. Finally, in Section 4.3, we present and discuss our findings related to the first research question, before transitioning to our second research question.

### 4.1 Methodology

The primary objective of this work is to explore the feasibility of applying the attention biasing knob (Hazarika et al. [2021]) to the hotel domain in order to enhance and control the informativeness of the generated output. To this end, our first research question (RQ1) is as follows: To what extent can cross-attention biasing be employed to generate more informative review responses given textual descriptions and details about a hotel? In addressing this research question, we investigate the potential of applying attention biasing in two scenarios:

- **Experiment 1:** A model fine-tuned on all available review response pairs with additional knowledge snippets, totaling approximately 2 million training examples.
- **Experiment 2:** A model fine-tuned on a ‘high-quality’ subset of review-response pairs, which is identified using the filtering approach described in Kew and Volk [2022].

We refer to the model in the first experiment as **OD** (Original Data), and the one in the second experiment as **FD** (Filtered Data) for consistency throughout our thesis.

For Experiment 1 (OD), we employ a straightforward approach. We fine-tune the BART model with our hotel review response training data, appending knowledge texts as additional input. During the model’s inference, we apply the attention bi-

asing knob in the encoder-decoder attention, aiming to enhance the informativeness of the generated output. For Experiment 2 (FD), we fine-tune the BART model again, but this time, we use a filtered training dataset, representing 40% of the original dataset. The attention biasing knob is applied as in Experiment 1 (OD). The methodologies for attention biasing and filtering will be introduced in subsequent sections.

### 4.1.1 Cross-Attention Biasing

Cross-attention biasing adjusts the attention distribution, thereby varying the focus allocated to different parts of the input. The underlying assumption is that increasing focus on a specific part of the input will lead to the incorporation of more content from that section in the generated response. In the context of knowledge-enhanced generation, which aims to yield more grounded responses, attention biasing can be employed to reweight attention values towards knowledge inputs. Consequently, the generated responses incorporate or reflect the knowledge, thereby enhancing their informativeness.

In more detail, a bias vector  $\mathbf{b}_t$ , consisting of bias values for each input token, is newly introduced in the attention calculation. The original attention distribution is adjusted through element-wise multiplication with  $\mathbf{b}_t$ , creating a biased attention distribution (refer to Equation 2.2). This adjustment process is uniformly applied across multiple attention layer heads. The new biased attention distribution is then used to determine the attention allocated to each input token (more information can be found in Section 2.5).

Note that the bias vector  $\mathbf{b}_t$  is not a learned parameter, but rather, the values are manually set. For instance, in the original paper, bias values of 2, 5, 10, and 50 were explored. Higher values allocate stronger bias to the corresponding input.

We implemented this cross-attention biasing knob during inference on our hotel dataset to explore the applicability of this methodology in the review response generation task for the hotel domain.

### 4.1.2 Data Filtering

One of the most effective methods to enhance a model’s performance in achieving a particular objective is to incorporate more training data that exhibit the desired characteristics and/or reduce undesirable data (Kew and Volk [2022]). For example,

one of the challenges of review response generation models is that they tend to produce safe or generic responses, as they are prevalent in the training data and easy for the model to learn. Thus, to develop models that generate more specific outputs, minimizing the negative influence of generic data can help.

Kew and Volk [2022] investigated the task of generating more specific responses for online reviews in the hospitality domain by filtering out generic data. To measure the genericness of responses, they employed various filtering methods such as lexical frequency, semantic similarity compared to generic example sentences, and perplexity scores. By utilizing these methods, they identified and filtered out generic responses from the training data and then fine-tuned the model with the best, less generic 40% of training data examples<sup>5</sup>. Interestingly, even with a 60% reduction in the training data, the filtering process led to the improvement of the model performance with regard to generating more specific and useful responses.

Inspired by their work, we hypothesized that fine-tuning the model with ‘high-quality’ filtered data would yield a more suitable model. This model (FD), we expect, would demonstrate enhanced specificity and informativeness in its generated responses, compared to the first model (OD) in Experiment 1.

For our Experiment 2, we employed the perplexity filtering method to filter our hotel training data. The findings from both automatic and human evaluations from their study indicated that the perplexity filtering method achieved the best overall performance<sup>6</sup>. The perplexity filtering method assesses a response text at the document level by evaluating the perplexity of each response using a language model. Lower perplexity signifies less surprisal, while higher perplexity may indicate a high degree of noise and potentially ungrammatical text. Therefore, responses with mid-range scores are considered more specific compared to the two extremes.

We calculated the perplexity of every response in the training data using the fine-tuned distilled-GPT2 model on a hotel dataset. Example 4.1 displays the responses with the lowest and highest perplexity in our original training data. The scores are presented in parentheses at the end of each sentence. We were able to confirm the argument in Kew and Volk [2022] that sentences with the lowest perplexity are quite generic with less surprisal and can be applied to almost all reviews. On the other hand, the responses with the highest perplexity contain grammatical errors and noise. As argued in the reference study, the responses with the lowest perplexity

---

<sup>5</sup>The code can be found in this repository: [https://github.com/ZurichNLP/specific\\_hospo\\_respo](https://github.com/ZurichNLP/specific_hospo_respo) We implemented the filtering method following the examples provided in this repository.

<sup>6</sup>Yet, the top-scoring models with the filtered data demonstrated a poorer degree of diversity than human-generated responses. One possible explanation is that neural NLG models typically tend to generate higher-frequency words (Holtzman et al. [2019]).

will thus lead the model to generate generic responses or, in other words, discourage the model from generating more informative and specific responses even with the increased bias values. Conversely, responses with the highest perplexity will introduce noise into the model, resulting in decreased training quality.

**Example 4.1 — Perplexity of responses (before filtering)**

**Response with the lowest perplexity:**

- “we are delighted to hear you enjoyed your stay with us and we hope to welcome you back soon” (PPL 2.9)
- “thank you for taking the time to review our hotel and we look forward to your next visit” (PPL 3.3)
- “thank you for taking the time to review your stay with us” (PPL 3.4)

**Response with the highest perplexity:**

- “sorry for in convince happen” (PPL 52,110)
- “thank you mystery ta member .” (PPL 19,920)
- “... with ammmaaazzziinng guests <NAME> !” (PPL 16,490)

Upon evaluating the response quality across varying perplexity values, we opted to filter 40% of the training dataset, retaining the responses with perplexity scores corresponding to the 30th to 70th quantiles of the overall perplexity score distribution of the original data. Table 4 presents the distribution of perplexity scores before and after filtering. In the original data, perplexity scores ranged from 2.9 to 52,110. The distribution of the original data exhibits a right-skewed pattern, with a relatively small interquartile range compared to the overall range of the data. This implies that a substantial portion of the data is concentrated within a narrower interval. In contrast, the filtered data has a range spanning from 23.9 to 44.1.

Example 4.2 presents responses with the lowest and highest perplexity after filtering. The two responses do not reveal noticeable differences in terms of specificity. In relation to response length, the Pearson correlation coefficient between scores and response length is less than 0.01, indicating negligible correlation. A subsequent qualitative analysis of the samples with the lowest and highest perplexity ranges confirmed no discernible differences not only in specificity but also in response length. To address the second research question, we fine-tuned a model using the training data that was filtered utilizing the perplexity filtering methodology described previously.



	Original Training Data (100%)	Filtered Training Data (40%)
<b># of data</b>	2,248,217	899,287
<b>mean</b>	43.6	32.7
<b>std</b>	110.7	5.7
<b>min</b>	2.9	23.9
<b>PPL 25%</b>	22.0	27.7
<b>50%</b>	32.1	32.1
<b>75%</b>	48.6	37.3
<b>max</b>	52,110	44.1

Table 4: Distribution of perplexity scores before and after filtering. Note that the data points with empty responses are excluded from the statistics, thus the number of the original training data (100%) in this table differs from that described in the dataset section (Section 4.2.1).

#### Example 4.2 — Perplexity of filtered responses

##### Response with the lowest perplexity:

“ <GREETING> thank you for taking the time to share your experience with us and the <NAME> community . we could n’t agree with you more that our prime location allows our guests to explore and enjoy the many sights and sounds our beautiful city has to offer , located within walking distance from shops , restaurants and attractions . rest assured your valued feedback is duly noted and will be raised regarding the tv channels . <NAME> , we do hope to welcome you back again in the near future . <SALUTATION> ”(PPL 24)

##### Response with the highest perplexity:

“ <GREETING> thank you very much for taking the time to post a review . we are glad you enjoyed our excellent location , price and the great quality of our beds . here at tune hotels we strive to offer a good night sleep at affordable prices without compromising quality or customer service . regarding noise levels , we have plenty of rooms available situated at the back of the building , and on the upper floors away from the lively <LOC> night life which can be requested upon check - in , subject to availability . we hope to see you again in the future <SALUTATION> ”(PPL 44)

## 4.2 Experimental Setup

### 4.2.1 Dataset

The hotel review response dataset used in our experiments was obtained from TripAdvisor<sup>7</sup>, as part of the ReAdviser project<sup>8</sup>. This dataset comprises hotel reviews, human responses, descriptions and list of amenities along with some related meta information such as review ratings. The dataset is split into a training set (2,249,330 instances), a validation set (11,361 instances), and a test set (11,360 instances). Over 50% of the ratings in the dataset are 5-star ratings. The dataset covers six countries: the United States, the United Kingdom, Australia, Canada, Ireland, and New Zealand, and contains 8,715 unique establishments.

The dataset has undergone partial preprocessing, with tokenization already completed. In both the reviews and responses, certain named entities are masked with special tokens, such as URLs, names, locations, numbers, and emails using the spaCy library.

Example 4.3 shows an instance of our dataset. The review consists of a title and body, separated by the token *-SEP-*. The response typically begins with a greeting and thanks, such as “*Dear <NAME>, thank you for your review*”, and concludes with a salutation such as “*best regards, manager <NAME>*”. Amenities and descriptions provide background information on establishments, which serve as grounding text in our experiments. The amenities section lists the available amenities at the establishment, separated by commas. In contrast, the description section comprises free text consisting of phrases or full sentences.

---

<sup>7</sup><https://www.tripadvisor.com/>

<sup>8</sup><https://www.cl.uzh.ch/en/texttechnologies/research/machine-learning/Response-Generation.html>

**Example 4.3 — An example from the hotel dataset**

**Review:** “the fabulous strand —SEP— it is an amazing experience for us every christmas . this is our third year as a family and the lengths they go to to make it a magical dream for our daughter is unbelievable . the standard of the <LOC> , quality and service is in a league of its own . this is where we will be going every christmas for as long as we can . thank you , to the strand”

**Response:** “<NAME> , thank you so much for your very kind words . its great to hear that you and your family continue to have such a lovely time spending christmas with us here at the limerick strand hotel , i ’m especially glad to hear that your daughter found it a magical time ! we really look forward to welcoming you and your family back in the very near future best regards , <NAME> deputy general manager”

**Amenities:** parking , free high speed internet ( wifi ) , pool , fitness centre with gym / workout room , bar / lounge , babysitting , children ’s television networks , car hire , wifi , indoor pool , heated pool , plunge pool , shallow end in pool , fitness / spa changing rooms , (...)

**Description:** an award winning 4\*hotel ideally located in the heart of limerick city centre . with stunning vistas of the river shannon and a birds - eye view of every landmark in the historical city the limerick strand hotel is the ideal venue for both business and pleasure trips . the hotel offers a friendly , inviting atmosphere where guests feel comfortable and well taken care of (...)

**4.2.2 Preprocessing and Input Formatting**

In addition to tokenization and named entity masking, we masked salutations in the responses. An existing sequence labeling model from the ReAdvisor project was employed for the salutation removal. This salutation removal model replaces salutation sequences of the responses with new tags <GREETING> and <SALUTATION>, utilizing Flair embeddings (Akbik et al. [2018]). The model achieved a macro accuracy of 90% for both tags. Example 4.4 demonstrates a pair of responses before and after applying the salutation removal model. The blue parts in the original response were masked in the salutation-removed response. The last two box plots in Figure 6 indicate that masking salutations reduced the response length, albeit not on a large scale, compared to the original response.

**Example 4.4 — An example of salutation removal****Original response:**

“**dear** <NAME> , thank you for taking the time to share your experience on tripadvisor . we appreciate the kind words regarding the ambiance and décor of the fireside inn . we hope to welcome you back again soon . **yours in** <LOC> , <NAME> , **front office manager**”

**Salutation-removed response:**

“<GREETING> thank you for taking the time to share your experience on tripadvisor . we appreciate the kind words regarding the ambiance and décor of the fireside inn . we hope to welcome you back again soon . <SALUTATION>”

For input formatting, we adapted the methodology from Hazarika et al. [2021], described in Section 3.1.2. In contrast to the reproduction study, here our input buckets correspond to the review, amenities, and description. The bucket size for the review was fixed at 200 tokens. As additional knowledge snippets, 200 tokens were assigned for amenities, and 240 tokens for the description. Figure 6 displays the token length percentiles of each input bucket in the entire dataset. As in the original study, the assigned bucket sizes approximately correspond to the 75th percentile of the dataset. Again, in instances where a bucket’s token count exceeded the defined size, the text within that bucket was truncated. Conversely, if a bucket held fewer tokens than its set size, the bucket was padded to the bucket size on the right. As a result, each input length was fixed at 652 (amenities (200) + description (240) + review (200) + special tokens (12)). Refer to Example 4.5 for the placement of special tokens (in blue) in the input. To reduce noise during training, we processed and fed the input only when all the sections (review, response, amenities, and description) were not empty strings.

**Example 4.5 — Input formatting for our experiment**

<s> <amenities> amenities <description> description <review> review </s>

**4.2.3 Fine-tuning and Inference**

We closely adhered to the experimental setup used in the preliminary experiment (Chapter 3). We fine-tuned the BART<sub>BASE</sub> model on our dataset. Since the hotel dataset volume is substantially larger than the topical dataset used in the preliminary experiment, we adjusted the training epochs to 2, with a total batch size of

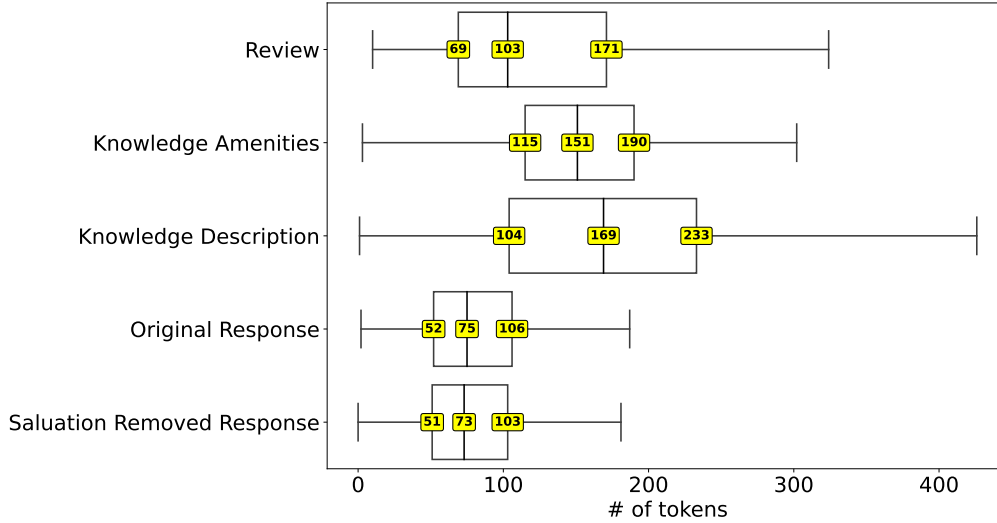


Figure 6: Distribution of token length in the dataset, including the original train, validation, and test sets. The numbers in yellow represent the token lengths at the 25th, 50th, and 75th quantiles, respectively. Outliers are excluded for this statistic.

40, and employed the same learning rate of  $6.25e-5$ . Note again that the attention biasing knob is not applied during training.

For inference, we again utilized nucleus sampling with a top-p value of 0.9, and the temperature was set at 1.0.<sup>9</sup> The maximum generation length was set to 256. While this value far exceeds the 75th percentile of the dataset (103), we chose a longer length to account for the possibility that more informative knowledge-enhanced responses might be longer than the gold response. We apply the cross-attention biasing knob during inference and experimented with four biasing profiles (**Review**, **Amenities**, **Description**, **Knowledge**). The profiles were configured as follows:

**Review:** In the Review profile, the knowledge bias (both amenities and description) was fixed at 1, while a higher bias value was assigned to the review part. Specifically, we set the bias values to  $(b_t^a, b_t^d, b_t^r) = (1, 1, 5)$  for all timesteps, where  $b_t^a$ ,  $b_t^d$ ,  $b_t^r$  mean amenities bias, description bias and review bias, respectively.

<sup>9</sup>The original temperature setup in Hazarika et al. [2021]’s experiment (0.7) produced texts with a high degree of degeneration and non-sensical text in our preliminary inference analysis with a smaller dataset. The majority of the generated outputs contained hallucinations in this setting. Consequently, we conducted our main experiments with the default temperature value of 1.0.

**Amenities:** In the Amenities biasing profile, the description and the review bias were fixed at 1, and an increased value was given to the amenities bias. We experimented with various amenities bias values to examine the level of control as the bias value escalated. Specifically, we set  $(b_t^a, b_t^d, b_t^r) = (\delta, 1, 1)$  where  $\delta \in \{2, 5, 10, 50\}$  for all timesteps. Note that higher values ( $b_t^a$ ) allocate stronger bias to the amenities part.

**Description:** In the Description biasing profile, the amenities and the review bias were fixed at 1, and an increased value was given to the description bias. Specifically, we set  $(b_t^a, b_t^d, b_t^r) = (1, \delta, 1)$  where  $\delta \in \{2, 5, 10, 50\}$  for all timesteps.

**Knowledge:** In the Knowledge biasing profile, the review bias was fixed at 1, and an increased value was given to the knowledge bias with the same amount (amenities bias and description bias). Specifically, we set  $(b_t^a, b_t^d, b_t^r) = (\delta, \delta, 1)$  where  $\delta \in \{2, 5, 10, 50\}$  for all timesteps.

**No Bias (Baseline):** The baseline without attention biasing was set to  $(b_t^a, b_t^d, b_t^r) = (1, 1, 1)$  for all timesteps.

## 4.2.4 Setup for Experiment 2

For our Experiment 2 (the FD model), we fine-tuned the  $\text{BART}_{\text{BASE}}$  model using filtered training data, which corresponds to 40% of the entire training data. Validation was also performed on the filtered dataset, employing the same PPL filtering methodology. We trained the model for three epochs. All the other experimental setups are identical to Experiment 1 (the OD model) described in Sections 4.2.2 and 4.2.3.

## 4.2.5 Evaluation

We assess our models in terms of informativeness, diversity, alignment with gold human responses, and response perplexity. Due to resource limitations, we do not conduct human evaluations. To partially address this limitation, and as our primary objective is to enhance the informativeness of generated responses, we employ three distinct types of metrics for assessing informativeness.

**Informativeness:** For our purpose, we define ‘informativeness’ as the degree to which the details from the provided knowledge snippet are integrated into the generated responses, in line with our preliminary experiment. To this end, we employ  $\text{ChrF}++_k$ ,  $\text{BERTScore}_k$ ,  $\text{unigram precision}_k$  and  $\text{recall}_k$ . While these

reference-based metrics are not traditionally indicators of informativeness, we draw inspiration from the evaluation methodology in Hazarika et al. [2021] and calculate these metrics with respect to the concatenated knowledge snippets (amenities and description). This approach enables us to assess the extent of lexical or semantic overlap between the generated response and the knowledge parts, thereby providing a measure of informativeness.

**Diversity:** We postulate a relationship between informativeness and diversity, presuming that a model grounded in the knowledge component would generate not only more informative text but also exhibit increased inter-textual diversity. To evaluate diversity, we employ **Self-BLEU**.

**Alignment with gold responses:** We assess the alignment with gold responses as a supplementary assessment. For this purpose, we employ **ChrF++<sub>tg</sub>**<sup>10</sup> with respect to gold responses. This metric assists in indirectly evaluating the contextual appropriateness of the generated text, i.e., whether the generated text adheres to the response scheme of the human responses.

**Perplexity:** We calculate **perplexity** scores using the Distilled-GPT2 model. In the preliminary experiment to replicate Hazarika et al. [2021]’s work, we used perplexity as a measure of fluency, in accordance with the original study’s evaluation methodology. However, we also consider another perspective on perplexity in our main experiments. Higher perplexity indicates that the model finds the text less predictable, which could be due to several factors. For instance, it may be a result of using less frequent words or forming longer sentences (Jin et al. [2022]). Such factors can be linked to specificity, suggesting that higher perplexity could denote less generic text (also observed in the data filtering in Section 4.1.2). Therefore, if the increase in perplexity is accompanied by enhanced informativeness and diversity, we cautiously interpret it as a proxy for specificity.

## 4.3 Results

In this section, we present the results of our two experiments for our first research question, compare their outcomes, and discuss potential enhancements. As a re-

---

<sup>10</sup>Please note that ChrF++ is also employed in the evaluation of informativeness. To distinguish between these two applications, we have added a subscript. The subscript ‘tg’ indicates that the metric is calculated with respect to the ‘target’ (human response), thus evaluating alignment with gold responses. Conversely, the subscript ‘k’ denotes ‘knowledge’, meaning that the metrics are calculated with the knowledge snippet (amenities and description concatenated) as a reference for assessing informativeness.

minder, we refer to the model in Experiment 1 as OD and the one in Experiment 2 as FD.

### 4.3.1 Experiment 1: Fine-tuning with Original Data (OD)

Experiment 1 adopts a simple approach, adhering closely to the original study by Hazarika et al. [2021]. The model is fine-tuned using the entire training dataset (about 2 million examples), with cross-attention biasing applied on different parts of the input during inference. Table 5 provides a summary of the results for Experiment 1.

Results for Experiment 1 (OD)

	<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
	<b>PPL</b>	<b>Self-BLEU</b> ↓	<b>ChrF<sub>lg</sub></b>	<b>ChrF<sub>k</sub></b>	<b>BERTScore<sub>k</sub></b>	<b>Precision<sub>k</sub></b>	<b>Recall<sub>k</sub></b>
<b>Baseline</b>	65.9	72.7	37.4	7.2	77.5	28.5	5.5
<b>Review 5</b>	62.1	70.4	39.7	8.3	77.6	28	6.4
<b>Amenities 2</b>	66.5	72.9	36.7	6.9	77.4	28.5	5.3
<b>Amenities 5</b>	65.8	73.5	35.6	6.5	77.4	28.5	5
<b>Amenities 10</b>	64.4	74.6	34.7	6.1	77.4	28.6	4.8
<b>Amenities 50</b>	59.1	77.9	32.2	5.5	77.3	28.6	4.3
<b>Description 2</b>	67.0	72.8	37.3	7.2	77.5	28.8	5.6
<b>Description 5</b>	67.2	73.2	36.9	7.2	77.5	29.1	5.6
<b>Description 10</b>	66.6	73.6	36.4	7.1	77.5	29.4	5.6
<b>Description 50</b>	67.4	74.4	35.3	7.0	77.6	29.9	5.6
<b>Knowledge 2</b>	57.3	72.7	36.7	6.9	77.5	28.7	5.4
<b>Knowledge 5</b>	58.7	72.8	35.8	6.7	77.4	28.9	5.2
<b>Knowledge 10</b>	69.0	73.0	35.3	6.5	77.4	28.9	5.1
<b>Knowledge 50</b>	70.2	73.3	34.7	6.4	77.4	29.1	5

Table 5: Evaluation metrics for Experiment 1 (OD). Note that ChrF refers to ChrF++ and BERTScore denotes the F1 score. The reported results are the average across five random seeds. The standard deviation for these results can be found in Table 10 in the Appendix.

#### Informativeness:

To gauge the influence of attention biasing on informativeness, we compare the results to the baseline. In the table, green signifies an increase in score compared to the baseline, while red represents a decline. For the Self-BLEU, where lower is better, a decrease is marked in green. Darker colors indicate a higher degree of gain or loss.



For the Review 5 profile, where the bias value of 5 is assigned to the review and 1 to the knowledge part, a slight increase in the informativeness metrics for  $\text{ChrF}_k$ ,  $\text{BERTScore}_k$ , and  $\text{Recall}_k$  is observed, although  $\text{precision}_k$  drops. For the knowledge category profiles<sup>11</sup> (i.e. Amenities N, Description N, and Knowledge N), where greater attention is given to the knowledge part, there is a general decrease in  $\text{ChrF}_k$  as the biasing value increases.  $\text{Recall}_k$  also decreases in the case of Amenities and Knowledge profiles. The Description profile shows an increase in  $\text{recall}_k$ , but the improvement is minimal, with only a +0.1 increase compared to the baseline for every bias value.

Interestingly, the Review profile shows increased informativeness, while the knowledge category profiles exhibit a decrease. This contradicts our expectation that the knowledge category profiles would integrate more from the knowledge snippets, thus improving informativeness.

Another notable aspect is that  $\text{precision}_k$  consistently rises with the bias value in all knowledge category profiles, a distinct behavior that will be discussed further in Section 6.1.1.

In brief, the Review 5 profile enhances informativeness, whereas the knowledge category profiles do not show the anticipated results. Moreover, the OD model does not exhibit the desired controllability (i.e., whether increased profiles lead to higher informativeness). These two observations indicate that attention biasing in the OD model does not effectively enhance or control the informativeness of responses.

### **Diversity:**

Self-BLEU increases as more attention is given to the knowledge part compared to the baseline, suggesting that more attention biasing leads to less diverse output. This is also counterintuitive, as more incorporation from different knowledge texts is expected to lead to more diverse output. This observation indirectly suggests that the desired increase in informativeness is not achieved when more bias is applied.

### **Alignment with gold responses:**

A decrease in  $\text{ChrF}_{tg}$  is observed when more bias values are assigned in all profiles, compared to the baseline, except for the Review profile. This indicates that the generated output becomes less similar to the human gold responses. If this was

---

<sup>11</sup>To eliminate ambiguity, we use the term ‘knowledge category profiles’ to encompass the Amenities, Description, and Knowledge profiles. However, when referring exclusively to the Knowledge profile, we always use capitalization, as in ‘Knowledge profile’

accompanied by an increase in informativeness, it could be explained by the fact that the system is incorporating more knowledge, and thus deviating from the gold responses. The simultaneous decrease in gold alignment and informativeness when attention biasing is applied to the OD model points to a potential decline in the overall generation quality.

### **Perplexity:**

The perplexity displays a fluctuating pattern when more bias is applied. While the Description profile generally exhibits a marginal increase in perplexity compared to the baseline, the Amenities and Knowledge profiles show both improvements and deteriorations. The lack of a consistent trend of the OD model raises questions about the stability of the generated content in terms of fluency and specificity.

### **Overall Results:**

The key findings from the experiment with the OD model are as follows: The attention biasing does not effectively enhance or control the informativeness of the generated responses with the knowledge category profiles. Moreover, decreases in diversity (as measured by Self-BLEU) and gold alignment ( $\text{ChrF}_{tg}$ ), accompanied by fluctuating perplexity indirectly suggest that attention biasing applied to this model also diminishes the overall inference performance, leading to a degradation in the quality of the text.

## **4.3.2 Experiment 2: Fine-tuning with Filtered Data (FD)**

Experiment 2 presents a slight modification to the fine-tuning approach deployed in the OD model, by utilizing the subset of training data filtered using perplexity threshold. This filtered data aims to produce outputs of increased specificity and informativeness by decreasing the genericness of the data. A summary of the results for Experiment 2 can be found in Table 6. Note that the scores for both the OD and FD models were obtained using the same test set to ensure a fair comparison.

### **Informativeness:**

For the Review 5 profile, there is a marginal rise in informativeness metrics for  $\text{ChrF}_k$  and  $\text{Recall}_k$ .  $\text{BERTScore}_k$  remains the same, whereas  $\text{Precision}_k$  dropped. For the

## Results for Experiment 2 (FD)

	<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
	<b>PPL</b>	<b>Self-BLEU</b> ↓	<b>ChrF<sub>tg</sub></b>	<b>ChrF<sub>k</sub></b>	<b>BERTScore<sub>k</sub></b>	<b>Precision<sub>k</sub></b>	<b>Recall<sub>k</sub></b>
<b>Baseline</b>	70.1	70.5	36.8	7.6	77.6	28.6	6
<b>Review 5</b>	66.8	68.5	38.5	8.6	77.6	27.9	6.8
<b>Amenities 2</b>	69.8	70.8	36.1	7.3	77.5	28.6	5.8
<b>Amenities 5</b>	69.0	71.6	35.2	6.9	77.5	28.7	5.5
<b>Amenities 10</b>	67.4	72.7	34.4	6.6	77.4	28.8	5.2
<b>Amenities 50</b>	60.3	75.7	32.4	6.0	77.3	28.7	4.8
<b>Description 2</b>	71.2	70.6	36.5	7.5	77.6	28.9	6
<b>Description 5</b>	72.3	71.0	35.7	7.4	77.6	29.4	5.9
<b>Description 10</b>	72.7	71.4	35.1	7.2	77.6	29.8	5.8
<b>Description 50</b>	72.4	72.4	33.3	6.7	77.5	30.3	5.4
<b>Knowledge 2</b>	71.5	70.7	36.0	7.3	77.6	28.9	5.8
<b>Knowledge 5</b>	62.4	71.1	35.1	7.0	77.5	29.2	5.5
<b>Knowledge 10</b>	73.0	71.5	34.5	6.8	77.5	29.4	5.4
<b>Knowledge 50</b>	73.5	71.9	33.6	6.6	77.5	29.7	5.3

Table 6: Evaluation metrics for Experiment 2 (FD). The reported results are the average across five seeds. The standard deviation for these results are reported in Table 11 in the Appendix.

profiles of knowledge categories, all metrics, barring precision<sub>k</sub>, depict a decrease in scores as more attention is given. These trends are similar to those observed in the OD model, challenging our initial expectations. The FD model does not exhibit the desired behavior as well, indicating that attention biasing does not effectively enhance or control the informativeness of responses.

### Diversity:

The trend observed in diversity aligns with those noted in the OD model. As more attention is directed towards the knowledge category 2 profiles, Self-BLEU increases, suggesting that the text generated with greater bias tends to be less diverse.

### Alignment with gold responses:

The results demonstrate a decrease in ChrF<sub>tg</sub> when more bias value is applied to knowledge. This pattern is consistent with that exhibited in the OD model. Again, the decline in both informativeness and gold alignment could potentially indicate a general decrease in the overall quality of the output.

**Perplexity:**

The perplexity scores in this model exhibit a degree of fluctuation. Amenities profile shows a decrease in perplexity as more bias value is given, while the scores in the Description and Knowledge profiles increase, with the exception of Knowledge 5.

**Overall Results:**

The patterns observed across informativeness, diversity, alignment with gold responses, and perplexity in the FD model closely mirror those of the OD model. These results indicate that the use of filtered data, despite its reduced genericness, did not augment the efficacy of the attention biasing knob in either controlling or enhancing the informativeness of the generated outputs.

These findings underscore the complexity inherent in effectively utilizing attention biasing and suggest that other factors beyond data genericness may play crucial roles in shaping the outputs. Further discussion in this regard will be presented in Section 4.3.4.

**4.3.3 Comparison between OD and FD models**

Table 7 provides a comparison of the scores achieved by the OD and FD models. Light green highlights where the FD model outperforms the OD model, while light red indicates instances where the FD model performs on par with or worse than the OD model.

A key observation is that the FD model generally achieves higher scores on informativeness metrics compared to the OD model. This suggests that the FD model is capable of generating more informative responses. Along with this, the FD model consistently records a lower Self-BLEU score, indicating that it generates more diverse output. These findings align with the results of the experiments conducted in Kew and Volk [2022], which suggested that models trained on filtered data display higher levels of diversity and specificity compared to models trained on the entire dataset. The FD model consistently exhibits slightly higher perplexity scores, which, when considered alongside the improved informativeness and diversity, further reinforces the idea that it is capable of generating less generic responses.

Score Differences of OD and FD

	<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
	<b>PPL</b>	<b>Self-BLEU</b> ↓	<b>ChrF</b> <sub>tg</sub>	<b>ChrF</b> <sub>k</sub>	<b>BERTScore</b> <sub>k</sub>	<b>Precision</b> <sub>k</sub>	<b>Recall</b> <sub>k</sub>
<b>Baseline</b>	4.2	-2.2	-0.6	0.4	0.1	0.1	0.5
<b>Review 5</b>	4.7	-1.9	-1.1	0.3	0	-0.1	0.4
<b>Amenities 2</b>	3.3	-2.1	-0.5	0.4	0.1	0.1	0.5
<b>Amenities 5</b>	3.2	-1.9	-0.4	0.4	0.1	0.2	0.5
<b>Amenities 10</b>	3.0	-2.0	-0.2	0.4	0	0.2	0.4
<b>Amenities 50</b>	1.2	-2.2	0.2	0.5	0	0.1	0.5
<b>Description 2</b>	4.2	-2.2	-0.8	0.3	0.1	0.1	0.4
<b>Description 5</b>	5.1	-2.2	-1.2	0.2	0.1	0.3	0.3
<b>Description 10</b>	6.1	-2.2	-1.3	0.1	0.1	0.4	0.2
<b>Description 50</b>	5.0	-2.0	-2.0	-0.3	-0.1	0.4	-0.2
<b>Knowledge 2</b>	14.3	-1.9	-0.7	0.4	0.1	0.2	0.4
<b>Knowledge 5</b>	3.6	-1.7	-0.8	0.3	0.1	0.3	0.3
<b>Knowledge 10</b>	4.0	-1.5	-0.8	0.3	0.1	0.5	0.3
<b>Knowledge 50</b>	3.3	-1.4	-1.1	0.2	0.1	0.6	0.3

Table 7: Score differences of OD and FD models. Light green highlights where the FD model outperforms the OD model, while light red indicates instances where the FD model performs on par with or worse than the OD model. The reported results are the average over five seeds.

### 4.3.4 Discussion

In our first research question, we discovered that attention biasing does not enhance or control the informativeness in the generated responses for our hotel review response dataset. This contrasts with the preliminary experiment on dialog generation tasks, which demonstrated that attention biasing could effectively enhance informativeness. A potential explanation for this discrepancy lies in the differences between the datasets used.

Figure 4 and Figure 6 illustrate substantial differences in the token number distribution for each input section. In the Topical Chat dataset, used in the preliminary experiment, the knowledge snippet usually consists of a single sentence, and the knowledge bucket size is fixed at 32. In contrast, our dataset features much longer sequences comprising multiple sentences in the knowledge sections, leading to bucket sizes fixed at 200 for amenities and 240 for descriptions. Furthermore, in the Topical Chat dataset, the knowledge snippets paired with the dialog turns are selected based on the highest similarity, which would encourage the model to learn groundedness

in its generation. We hypothesize that these differences in the datasets led to the distinct behaviors observed.

Therefore, in Section 5, we continue our experiments under the second research question to identify a setting where attention biasing can be effectively applied. We especially focus on the input length of knowledge snippets and conduct inference experiments with varying knowledge bucket sizes.

# 5 RQ2 Experiment

This chapter carries forward our investigation into the effectiveness of attention biasing under different setups, addressing our second research question. Section 5.1 introduces the motivation behind this experiment. The modified experimental setup, in contrast to the setup used in the previous chapter, is described in Section 5.2. Lastly, Section 5.3 presents the results obtained from the experiment.

## 5.1 Motivation

As discussed in Section 4.3.4, there could be several potential reasons for the ineffectiveness of attention biasing when applying it to our task. One possible reason could be the discrepancy in the knowledge input length. Accordingly, we formulated our second research question as follows: How does the modification of knowledge input length during the inference stage impact the model’s performance with respect to informativeness?

Our working hypothesis posits that the length of the knowledge input sequence could potentially impact the model’s ability to enhance informativeness when employing attention biasing. Knowledge sequences that are too short might lack sufficient information to extract, even with adequate bias values. Insufficient information, coupled with a lack of context, could also result in incoherent outputs. On the other hand, an excessively long knowledge input length may cause the model to struggle to focus on the most relevant information, resulting in the dilution of the information.

Several studies have discussed the relationship between input length and model performance. Kew et al. [2020] benchmarked the work of Gao et al. [2019], which investigated a review response generation for the mobile app domain. This benchmarking study employed a seq2seq model with encoder-decoder attention, applying it to the hospitality domain. They found that the adapted approach encountered difficulties in generating suitable review responses when applied to the hospitality domain. They speculated that the discrepancies in input text lengths between the dataset in Gao et al. [2019] and their study might be one of the underlying causes

for comparatively low performance. Specifically, app reviews are generally much shorter in length than reviews for the hospitality domain.

Gao et al. [2021] conducted an analysis of sequence length for review response generation, expanding their previous study (Gao et al. [2019]). They utilized Gated Recurrent Unit (GRU)-based seq2seq models with attention mechanism (Cho et al. [2014]). They observed that the sequence lengths had an impact on model performance. For instance, a model with a review input length of 100 demonstrated a slight increase in BLEU score compared to a model with a review input length of 200.

Although these two studies employed different architectures than our current study, the influence of input length is a common concern across various models in NLG. Consequently, in addressing our second research question, we aim to investigate whether changes in knowledge input lengths could lead to different outcomes.

## 5.2 Experimental Setup

For this experiment, we employed the fine-tuned OD model from Experiment 1, making modifications to the input formatting during inference. We reduced the bucket sizes of amenities and description sections and conducted inference experiments with three distinct sizes: [40, 100, 160]. For instance, for a bucket size of 40, only the initial 40 tokens from amenities and description are fed as input, with any subsequent tokens truncated. Unlike in the OD and FD models where the knowledge input length was fixed at 200 for amenities and 240 for descriptions, this experiment ensured a uniform bucket size for both knowledge types. We experimented with Amenities, Description, and Knowledge profiles. Apart from these changes, the experimental setup remains consistent with the setup in Experiment 1 described in Sections 4.2.2 and 4.2.3.

## 5.3 Results

This section details the results of this experiment, employing automatic evaluation metrics supplemented with a qualitative example, and comparison between different settings. We term this particular experiment as Experiment 3 for ease of reference. The inference settings with distinct bucket sizes are labeled as KB40, KB100, and KB160, corresponding to Knowledge Bucket sizes of 40, 100, and 160, respectively.



### 5.3.1 Automatic Evaluation

Table 8 summarizes the automatic evaluation metrics for KB100. Given the similar trends across all three KB settings and in the interest of conciseness, we report the results for KB40 and KB160 in Table 12 and Table 14, respectively, in the Appendix. Any deviations in behavior among the three settings will be explicitly discussed in Section 5.3.2.

Results for Experiment 3 (KB100)

	<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
	<b>PPL</b>	<b>Self-BLEU</b> ↓	<b>ChrF</b> <sub>tg</sub>	<b>ChrF</b> <sub>k</sub>	<b>BERTScore</b> <sub>k</sub>	<b>Precision</b> <sub>k</sub>	<b>Recall</b> <sub>k</sub>
<b>Baseline</b>	64.4	86.6	28.8	5.1	77.2	28.6	4.1
<b>Amenities 2</b>	65.2	86.7	27.9	4.8	77.2	28.5	3.9
<b>Amenities 5</b>	67.0	87.0	26.6	4.5	77.2	28.4	3.6
<b>Amenities 10</b>	67.4	87.6	25.6	4.2	77.1	28.4	3.4
<b>Amenities 50</b>	69.1	88.1	23.8	3.8	77	28.3	3.2
<b>Description 2</b>	68.0	86.5	29.1	5.2	77.3	29	4.2
<b>Description 5</b>	71.9	86.5	29.3	5.4	77.4	29.7	4.4
<b>Description 10</b>	74.6	86.3	29.3	5.5	77.4	30.2	4.5
<b>Description 50</b>	78.0	85.8	29.0	5.6	77.5	31	4.6
<b>Knowledge 2</b>	68.7	86.5	28.3	5.0	77.3	28.9	4
<b>Knowledge 5</b>	74.3	86.4	27.7	4.8	77.3	29.3	3.9
<b>Knowledge 10</b>	78.0	86.2	27.4	4.8	77.3	29.4	3.9
<b>Knowledge 50</b>	83.3	85.8	27.0	4.7	77.3	29.6	3.9

Table 8: Evaluation metrics for the KB100 setting. Note again that ChrF refers to ChrF++ and BERTScore denotes the F1 score. The results are averaged across five random seeds. The standard deviation for these results is reported in Table 13 in Appendix.

#### **Informativeness:**

The Amenities profile generally shows a decrease or no improvement in scores as the bias value increases. An exception is observed in the KB160 setting where precision<sub>k</sub> shows a slight increase (0.1-0.2) with more bias. This pattern, where all scores drop or remain static except for precision, mirrors the trend we see in the OD and FD models.

Contrarily, the Description profile reveals an improvement in all informativeness metrics across all settings (KB40, KB100, and KB160) as more bias is allocated to the description text. This demonstrates that the attention biasing mechanism functions as intended, integrating more content from the input as the bias values

increase. With regard to controllability, it is evident that the degree of informativeness can be steered by adjusting the bias values for the Description profile, as the scores increase monotonically with the bias value. The statistical significance of this improvement of  $\text{recall}_k$  was confirmed through a Wilcoxon signed-rank and a paired t-test. In all settings, we observed a p-value less than 0.05 compared to the baseline scores.

On the other hand, the Knowledge profile, which incorporates both amenities and description, yields mixed results. While  $\text{precision}_k$  and  $\text{BERTScore}_k$  show slight improvements, there are decreases in  $\text{ChrF}_k$  and  $\text{recall}_k$ . As a result, the Knowledge profile does not show a general improvement in informativeness.

To further illustrate the influence of bias values on the model’s informativeness with the Description profile, we provide a representative qualitative example that showcases common characteristics of the responses generated with our models. Example 5.2 presents responses generated under the KB100 setting with varying levels of bias in the Description profile. The corresponding review and description text are shown in Example 5.1.

### Example 5.1 — An example of a review and description

#### Review:

“well positioned , nice clean and quiet room, friendly staff , good price , what more can you ask ? —SEP— very happy with the room and features , small problem with the tv remote immediately fixed upon drawing to the attention of staff , how would they know otherwise . the quest bondi junction is conveniently located in close proximity to the train station and public transport with lot of shopping and food outlets nearby . free wifi , higher speed available at additional cost . would definitely stay again on the next trip down .”

#### Description (100 tokens):

quest bondi junction provides contemporary accommodation that delivers the excellent standards of service and guest experiences you have come to enjoy and expect from quest serviced apartments. the 82 spacious studio apartments are well equipped and stylishly furnished to satisfy the needs of the corporate and leisure traveller . quest bondi junction has an anticipated  $4\frac{1}{2}$  star rating and is conveniently located within walking distance of westfield shopping centre, centennial park and major public transport networks ( rail and bus ). this cosmopolitan location provides easy (...)

**Example 5.2 — Responses by KB100 (Description profile)****Response with Description 5:**

“<GREETING> thank you for taking the time to review your recent stay with us at **quest bondi junction**. we value all of our guest feedback as it allows us the opportunity to reflect on our services and facilities. i am very pleased to hear that you enjoyed your stay. we look forward to welcoming you back to <LOC> in the near future. <SALUTATION>”

**Response with Description 10:**

“<GREETING> thank you for taking the time to review your recent stay with us at **quest bondi junction**. we value all of our guest feedback as it allows us the opportunity to reflect on our services and facilities so they may better meet our guest needs. i am very pleased to hear that you enjoyed your stay in <DIGIT> of our **spacious studio apartments**. our studio apartments are **well equipped to cater for both corporate and leisure travellers**. we look forward to welcoming you back to <LOC> in the near future. <SALUTATION>”

**Response with Description 50:**

“<GREETING> thank you for taking the time to review your recent stay with us at **quest bondi junction**. we value all of our guest feedback as it allows us to continuously review and improve our service and facilities. i am so pleased to hear that you enjoyed your stay and the apartment met your needs and expectations. as you mentioned, our studio apartments are **well equipped to satisfy both the corporate and leisure traveller**. they are furnished to suit the needs of our guests. the apartments are **located in the heart of <LOC>, with easy access to <NAME> shopping centre, centennial park** and **the transport network**. we look forward to welcoming you back to the hotel in the near future. <SALUTATION>”

The highlighted segments in the descriptions and responses depict the information incorporated from the description into the model’s output. The difference in the responses becomes more pronounced as the bias factor increases. At Description 5, the response only refers to the name of the establishment, while at Description 10, the response also references the *spacious studio apartments* and their suitability for *corporate and leisure travelers*. At Description 50, the response incorporates more detailed information, including the location of the apartments relative to local amenities. This clearly demonstrates the increasing informativeness of the model’s responses as the bias value escalates.

Given that the Description profile can enhance and manage the informativeness of the responses, we focus on analyzing the remaining measurements of this profile to understand the impact of informativeness enhancement.

**Diversity:**

Under the KB100 settings, increasing the bias factor on the description text results in a decrease in Self-BLEU scores, suggesting that biasing the description text generates more diverse responses. This trend supports our hypothesis that there is a positive correlation between informativeness and diversity.

**Alignment with gold responses:**

For the Description profile,  $\text{ChrF}_{tg}$  initially rises with smaller bias values, but then falls relative to either the baseline or the preceding bias value as the bias value becomes larger. For instance, in the KB100 setting (as shown in Table 8), the scores rise up to Description 5, then show a decrease at Description 50. A similar pattern is observed in the KB40 setting, with a decrease in  $\text{ChrF}_{tg}$  at Description 50 compared to the baseline. The KB160 setting experiences a relatively larger drop in  $\text{ChrF}_{tg}$  at Description 50. Given that this drop in  $\text{ChrF}_{tg}$  for larger bias values coincides with an increase in informativeness, it can be interpreted that the model is incorporating more knowledge and consequently deviating from the gold responses, which tend to entail less knowledge.

**Perplexity:**

Across all KB settings, we observe an increase in perplexity as we apply larger bias values to description text. This pattern, coupled with a rise in informativeness and diversity, cautiously suggests that the model with KB settings is not necessarily decreasing fluency but rather generating responses that are less generic in nature.

### 5.3.2 Comparative Analysis of Inference Settings

Having established the efficacy of attention biasing with the Description profile, we now turn our focus to comparing the performance of the models with varying knowledge input lengths. For reference, we also include the results of the OD model. Note that all the results of the KB settings are derived from the same test set used for the OD and FD models.

When comparing these inference settings, it is critical to note that the reference-based metrics used ( $\text{ChrF}_{tg}$ ,  $\text{ChrF}_k$ ,  $\text{BERTScore}_k$ ,  $\text{precision}_k$ ,  $\text{recall}_k$ ) are computed with respect to knowledge snippets with original bucket sizes (200 for amenities and 240 for description). This means, for instance, a setting like KB100, which only considers the first 100 tokens of the amenities and description as input during inference, is still evaluated on longer amenities and description. While this approach ensures a more equitable evaluation, it generally results in lower scores for KB settings that process smaller knowledge snippets. Thus, our primary focus is on how these scores shift with increased bias, rather than their absolute values.

### Informativeness:

Figure 7 displays the variation in  $\text{recall}_k$  and effect size, as measured by Cohen’s  $D^{12}$ , across different bias values in the Description profile.

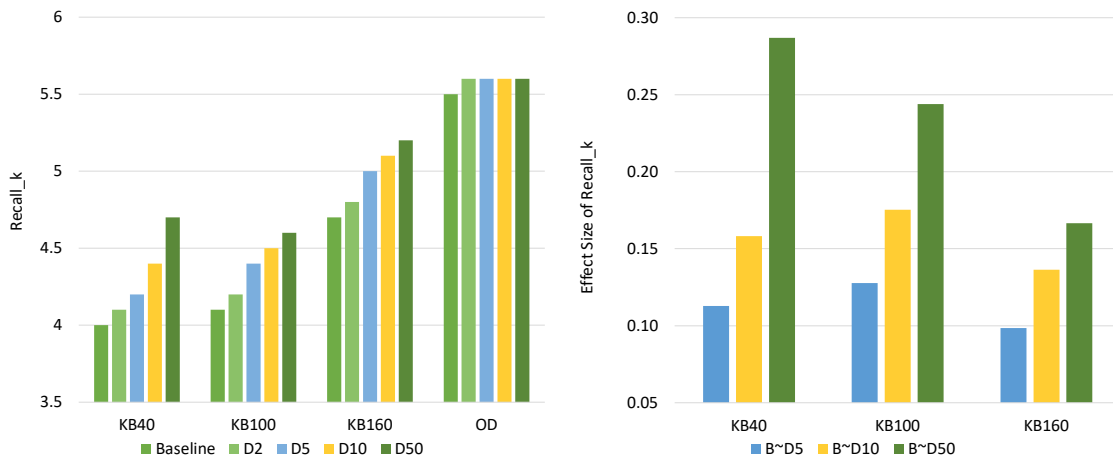


Figure 7: The graph on the left shows the variations in  $\text{recall}_k$ , while the graph on the right represents Cohen’s  $D$  effect size. In the legend, the term  $D$  refers to the Description profile, where, for instance,  $D5$  indicates Description 5. On the other hand,  $B$  represents the baseline. For example,  $B \sim D5$  signifies the effect size between the baseline and Description 5.

All KB settings demonstrate an increase in  $\text{recall}_k$  as bias intensifies, which is confirmed in the result analysis. It is important to clarify again that the absolute  $\text{recall}_k$  values differ between settings because the evaluation is based on longer text. Thus, KB40, which uses the least amount of knowledge, generally exhibits the lowest absolute scores, and KB160 the highest. To facilitate more comparative analysis, we also calculated the effect size for the KB settings. This metric progressively escalates

<sup>12</sup>Cohen’s  $D$  measures the magnitude of the difference between groups. The effect size  $d$  of 0.2, 0.5 and 0.8 correspond to small, medium and large effects (Cohen [1988]).

as the bias intensifies, indicating a more pronounced improvement in  $\text{recall}_k$  with a stronger attention bias towards the description.

Among the KB settings, the effect size of the KB160 setting tends to be consistently smaller compared to KB100 and KB40. This suggests that increasing the knowledge input size might result in smaller relative enhancements in  $\text{recall}_k$  when the attention bias is increased.

### Perplexity and Diversity:

Figure 8 presents the variation in perplexity and Self-BLEU across different settings as the bias value increases for the Description profile.

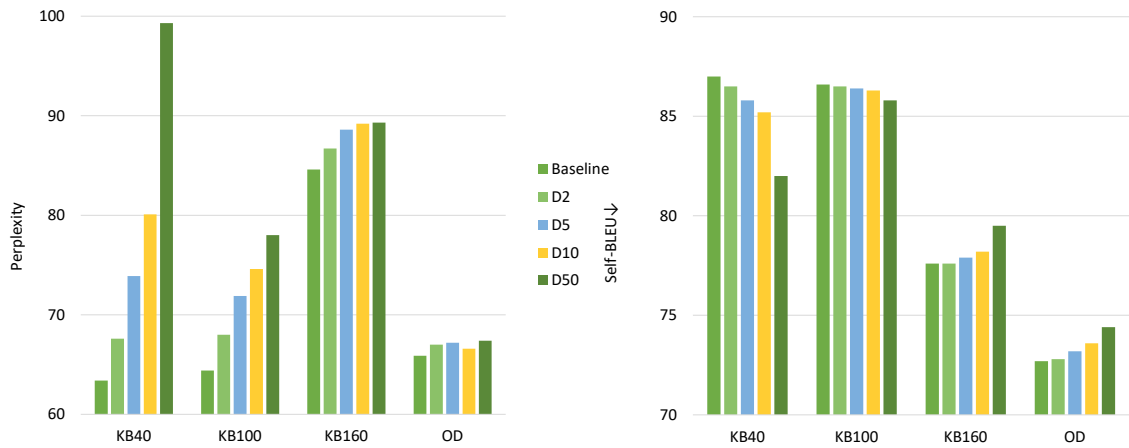


Figure 8: The graph on the left represents the variations in perplexity, while the graph on the right displays Self-BLEU.

In terms of perplexity, we observe an overall increase in all KB settings as more bias is introduced. Interestingly, this increase in perplexity is more prominent when the knowledge length is shorter, similar to the pattern observed in the effect size change (see Figure 7). Specifically, in the KB160 setting, there is a relatively modest increase in the effect size of informativeness and a relatively small increase in perplexity as well. This finding strengthens our argument for interpreting perplexity as an indicator of specificity rather than fluency in our task.

Furthermore, the average perplexity of human responses was measured to be 127, which is substantially higher than the perplexity of the texts generated by our models. This disparity can be attributed to the unique characteristics of human-generated text. Human responses naturally exhibit greater diversity, utilizing a wider vocabulary, a broader range of sentence structures, and encompassing a variety

of topics. These factors contribute to the increased unpredictability and higher perplexity observed in human-generated text. Therefore, given that our models still exhibit lower perplexity when compared to human responses, the slightly higher perplexity in our model’s outputs does not necessarily imply lower fluency, but rather reflects the incorporation of knowledge texts.

In terms of Self-BLEU scores, which are ideally lower, KB40 and KB100 show a decrease in scores as more bias is given. However, a different pattern is noticed in the KB160 setting, where an increase in bias for the Description profile actually leads to higher Self-BLEU scores, indicating a contrasting behavior, showcasing a contrasting behavior that aligns with the OD model. The examination of this divergent behavior under different knowledge input lengths will be further discussed in Section 6.1.2.

### **5.3.3 Key Findings**

In our key findings, we have observed that attention biasing effectively controls and enhances informativeness in all KB inference settings, which differs from the behavior observed in OD and FD models. This effectiveness, however, is only notable for the Description profile. It is noteworthy that the KB160 setting demonstrates trends more akin to the OD model in terms of diversity, perplexity, and gold alignment. From these observations, we infer the existence of an optimal knowledge type and input length that could lead to the effective application of attention biasing. In the following section, we will delve deeper into these results for a more detailed discussion.

## 6 Further Analysis and Discussion

This chapter delves into a thorough discussion of our research findings, beginning with a more in-depth analysis of the experimental results in Section 6.1. The subsequent Section 6.2 presents a comparative analysis with OpenAI’s ChatGPT, a state-of-the-art large language model, to highlight the unique characteristics and behavior of our model and provide insights for future improvement. Finally, in Section 6.3, we reflect on the limitations of our research and suggest potential directions for future work.

### 6.1 Decoding the Results: A Comprehensive Analysis of the Experimental Outcomes

In this section, we take a deeper dive into the intriguing results observed from our experiments. The goal is to interpret and explain these findings by examining the underlying mechanisms and potential factors influencing these outcomes. Specifically, we investigate the isolated increase in unigram precision<sub>k</sub> in the OD and FD models (Section 6.1.1), the role of knowledge input length in attention biasing efficacy (Section 6.1.2), and the differential impact of attention biasing on the Amenities profile (Section 6.1.3).

#### 6.1.1 Interpreting the Improvement of Precision<sub>k</sub> in OD and FD Models

As highlighted in Section 4.3, the OD and FD models showcased an intriguing pattern: the unigram precision<sub>k</sub> scores rose with increasing bias values across all profiles (Review, Amenities, Description, and Knowledge profiles). Interestingly, this trend did not correlate with enhancements in other measures of informativeness in the OD and FD models. This isolated increase in precision prompts further examination.



To understand this behavior, we first delve into the calculation of  $\text{precision}_k$ . Unigram precision is calculated as the ratio of common unigrams between the reference and prediction to the total number of unigrams in the prediction. Thus, an increase in precision can either be due to an increase in the number of common unigrams, or a decrease in the total number of unigrams in the prediction. Considering that other lexical-based metrics for informativeness such as  $\text{ChrF}_k$  and the unigram  $\text{recall}_k$  did not show corresponding improvement, we hypothesized that the increase in precision might be attributed to a reduction in output lengths.

In order to validate this hypothesis, we analyzed the average output length across all models and settings, as depicted in Figure 9. As anticipated, the OD and FD models, which did not demonstrate enhancements in informativeness under the application of attention biasing, displayed a noticeable reduction in average output length when more bias is allocated. Conversely, KB settings generally exhibited a slightly increased output length when more bias was given, compared to the baseline<sup>13</sup>.

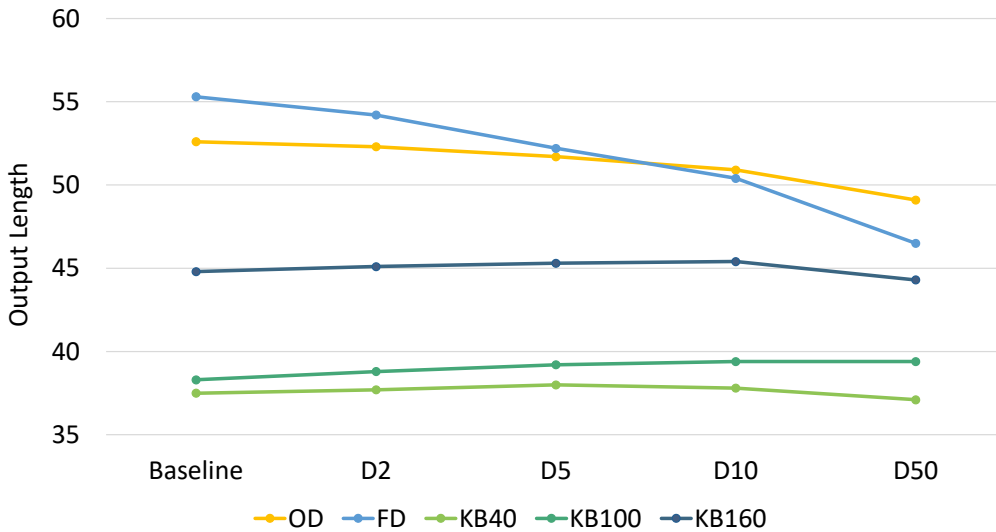


Figure 9: Average output length of different models and settings, calculated by word count separated by space. Results were averaged from five seeds.

Therefore, we can confirm that the increase in unigram  $\text{precision}_k$  in the OD and FD models primarily resulted from shorter outputs rather than reflecting a genuine

<sup>13</sup>We note that output length can serve as a proxy for model performance evaluation. As an illustration, Kew and Volk [2022] utilized output lengths as an evaluation metric, underpinning the notion that shorter responses typically signal a higher degree of genericness. Given this perspective, the tendency of KB settings to generate longer outputs with increased bias values can be interpreted as an indication of increased informativeness.

rise in the number of shared unigrams. Consequently, the observed escalation in  $\text{precision}_k$  in these models should not be misconstrued as an indication of improved informativeness.

### 6.1.2 Understanding the Role of Knowledge Input Length in Attention Biasing Efficacy

The contrasting effects of attention biasing across the FD, OD, and KB shed light on the crucial role of knowledge input length in determining the efficacy of attention biasing. The KB settings, characterized by their reduced knowledge input during inference, exhibited enhanced informativeness with increased bias. The FD and OD models, however, did not demonstrate the same pattern.

Intriguingly, variations were also observed within the KB settings. The KB160 setting, which showed improvement in informativeness with a knowledge length of 160, resulted in diversity, perplexity, and alignment to gold responses more similar to the OD model, where the knowledge lengths for amenities and description were fixed at 200 and 240, respectively. Also, the effect size of  $\text{recall}_k$  for the KB160 setting was the smallest among all KB settings.

These variations in model behavior underscore the critical role of the knowledge input length. The shortened knowledge potentially allows the attention mechanism to focus more effectively on the available knowledge, thereby generating more informative output when bias is applied towards the knowledge part.

Conversely, the FD and OD models grapple with a longer knowledge input, which might result in an information overload. This could challenge the attention mechanism's ability to identify and focus on the most pertinent knowledge, even with the introduction of biasing, resulting in shorter outputs without a corresponding increase in informativeness.

These findings align with prior research by Gao et al. [2021], which posited that sequence lengths impact the performance of a model with an attention mechanism. The study showed that a model with a review input length of 100 slightly outperformed a model with a review input length of 200 in terms of BLEU score. This suggests that longer knowledge inputs may not necessarily lead to better performance.

These insights collectively suggest that knowledge input length plays a pivotal role in determining the efficacy of attention biasing, carrying significant implications for

future model development. It emphasizes the necessity for careful consideration of knowledge input attributes when implementing attention biasing strategies.

### **6.1.3 Understanding the Differential Impact of Attention Biasing on the Amenities Profile**

One of the striking findings from our experiments is that attention biasing did not yield improvements for the Amenities profile, even within the KB settings. This contrasted with the effectiveness of biasing in the Description profile. To understand this, we explore the nature of the amenities and description texts (refer to Example 4.3 for an example of amenities and description).

Amenities, in our dataset, typically comprises a list of facilities or services available in the establishment. These features are discrete and factual, lacking narrative flow or contextual continuity. Unlike typical language use, amenities are essentially a collection of keywords, making it difficult to extract larger linguistic features beyond the word level. In contrast, descriptions provide a comprehensive overview of the establishment, using language in a more natural, narrative, and contextually integrated manner, akin to standard human conversation and writing.

The attention mechanism in transformer models, in their essence, is designed to identify and focus on relevant parts of the input, considering the interdependencies among tokens and capturing semantic and contextual relationships. However, in the case of the Amenities profile, the amenities texts lack sufficient contextual information. This could potentially impede the attention mechanism's ability to effectively integrate and extract meaningful and contextually relevant information from the amenities text.

This insight underscores the critical role that the nature of input data plays in the effectiveness of the attention mechanism. While attention biasing can improve model performance in scenarios with contextually rich and interconnected inputs (like descriptions), its impact can be limited for more fragmented and contextually poor inputs (like amenities). Future research could explore strategies to enhance the handling of context-poor inputs, as this type of information can easily be sourced and included as additional knowledge for review response generation.

## 6.2 Comparative Analysis with ChatGPT

With its launch in December 2022, OpenAI’s ChatGPT has quickly gained significant attention. Known for its impressive ability to generate human-like text, this model is widely recognized for its zero-shot, conversational interaction. The model’s remarkable advancement has led to extraordinary performances across a variety of NLP tasks such as reasoning and general text generation, which has ushered in unprecedented applications in numerous fields such as education, healthcare, human-machine interaction, scientific research, and even medicine (Liu et al. [2023]).

Given the widespread impact and prominence of ChatGPT, it serves as a compelling point of reference when evaluating new models and approaches. To this end, we conducted a comparative analysis of the method investigated in this thesis against ChatGPT. The primary focus of our analysis is qualitative, examining the responses generated by both our proposed model and ChatGPT, given identical input. This direct comparison reveals unique characteristics and capabilities inherent in each model, providing insights into their relative strengths and weaknesses.

For our analysis, we utilized the May 3 version of GPT-3.5-Turbo, commonly known as ChatGPT, and guided the model towards knowledge-enhanced generation. We sampled ten instances from our dataset, and instructed ChatGPT to generate responses while incorporating knowledge snippets at three levels: ‘low’, ‘moderate’, and ‘high’. This approach was designed to closely mirror the incremental bias value setup of our model. The appended knowledge snippets were either amenities, description, or a combination of amenities and description, approximating the biasing profiles used in our model. Nevertheless, note that this is a mere approximation due to the inherent prompt-based nature of ChatGPT, which does not allow for exact input formatting or parameter adjustment. An example prompt text used in our experiment is provided in Example 6.1.

**Example 6.1 — Prompt provided to ChatGPT (for grounding to the description)**

Please write three responses to the given review, grounded to the description, with varying degrees of groundedness (low, moderate, high).

Review: [review text]

Description: [description text]

## 6.2.1 Evaluation Metrics

While qualitative analysis serves as the main focus of our comparison, we supplement this with automatic metrics to provide a more objective evaluation and additional reference points. However, given the small sample size (10), these numbers should be taken with a grain of salt.

Table 9 summarizes the evaluation metrics for responses generated by ChatGPT. For ease of reference, we label responses grounded to the description as ‘Description profile’, responses grounded to amenities as ‘Amenities profile’, and responses grounded to both description and amenities as ‘Knowledge profile’. This terminology aligns with that used in our primary experiments.

		<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
		<b>PPL</b>	<b>Self-BLEU<sub>↓</sub></b>	<b>ChrF<sub>tg</sub></b>	<b>ChrF<sub>k</sub></b>	<b>BERT<sub>k</sub></b>	<b>Precision<sub>k</sub></b>	<b>Recall<sub>k</sub></b>
<b>Amenities</b>	Low	31.8	50.9	26.5	11.5	80.0	32.2	9.9
	Moderate	66.1	34.9	31.0	16.3	79.9	30.3	13.6
	High	48.9	27.9	31.6	21.6	80.0	28.0	18.1
<b>Description</b>	Low	37.6	42.1	30.3	13.4	79.7	28.5	10.7
	Moderate	87.0	30.8	32.4	19.2	80.0	29.9	15.4
	High	76.3	32.6	32.5	25.8	80.7	30.0	20.5
<b>Knowledge</b>	Low	28.6	41.0	30.6	13.5	79.6	29.2	10.6
	Moderate	71.6	34.0	31.2	17.1	79.6	27.1	13.4
	High	74.7	28.1	32.8	21.1	79.7	26.1	16.6

Table 9: Evaluation metrics for the responses generated by ChatGPT. Note that ChrF refers to ChrF++ and BERTScore denotes the F1 score.

Our analysis reveals that, similar to our models, ChatGPT improves substantially in informativeness across all metrics when the Description profile is used. However, a more intriguing observation arises in Amenities and Knowledge profiles. Unlike our KB settings, ChatGPT also demonstrates an increase in informativeness across these profiles: Both ChrF<sub>k</sub> and recall<sub>k</sub> metrics improve as the degree of groundedness increases. On the other hand, the precision<sub>k</sub> metric declines for these profiles.

This simultaneous increase in recall<sub>k</sub> and decrease in precision<sub>k</sub> suggests a specific pattern of output generated by ChatGPT. The increase in recall<sub>k</sub> indicates that ChatGPT incorporates a larger proportion of ‘informative’ tokens from the total set available in the knowledge text, as grounding gets stronger. However, the decrease in precision<sub>k</sub> underscores that while the model generates more ‘informative’ tokens, it also produces a larger volume of ‘non-informative’ tokens in the process. In other words, with the increase in groundedness in Amenities and Knowledge profiles,

ChatGPT’s output tends to be lengthier, but the fraction of the output that can be deemed informative or grounded decreases. These observations were further confirmed through a qualitative analysis, which we will elaborate on by providing specific examples in the next section.

As for the perplexity and Self-BLEU metrics, they show fluctuations across all profiles. Given that our evaluation is based on a small sample size of 10 instances, these fluctuations require further investigation to understand their implications.

## 6.2.2 Qualitative Analysis

In this section, we perform a close examination of specific instances to compare the performance of our proposed model (KB100 using the OD model) and the ChatGPT model across different profiles. Note that the characteristics of the responses we discuss are not confined to a single instance but are typically observed throughout the generated responses.

### 6.2.2.1 Description Profile Analysis

We commence our analysis by inspecting an instance grounded to the Description profile. Example 6.2 displays a set of review and description texts, with sections highlighted. The grey portions in the `description` indicate parts that our KB100 setting successfully incorporated into its response (please refer to Example 5.2 for the responses generated by KB100 for the same review and description instance). Lime segments highlight parts of the `review` that were integrated into the responses by ChatGPT. Yellow parts in the `description` are those that the ChatGPT model managed to incorporate, but our model missed.

Example 6.3 presents the responses generated by ChatGPT with low and high grounding, respectively, to the description in Example 6.2. As evident from the example, responses with higher groundedness clearly incorporate more text from the description. Moreover, compared to the responses generated by our model, the responses from ChatGPT tend to include more details from the review as well, addressing the issues mentioned therein. While the extent of this incorporation and whether it is optimal is still up for debate, this ability of ChatGPT is undeniably a positive attribute, potentially providing a direction for future improvements to our model.

**Example 6.2 — An example of a review and description****Review:**

“well positioned , nice clean and quiet room, friendly staff , good price , what more can you ask ? —SEP— very happy with the room and features , small problem with the tv remote immediately fixed upon drawing to the attention of staff , how would they know otherwise . the quest bondi junction is conveniently located in close proximity to the train station and public transport with lot of shopping and food outlets nearby . free wifi , higher speed available at additional cost . would definitely stay again on the next trip down .”

**Description:**

quest bondi junction provides contemporary accommodation that delivers the excellent standards of service and guest experiences you have come to enjoy and expect from quest serviced apartments. the 82 spacious studio apartments are well equipped and stylishly furnished to satisfy the needs of the corporate and leisure traveller . quest bondi junction has an anticipated 4½ star rating and is conveniently located within walking distance of westfield shopping centre, centennial park and major public transport networks ( rail and bus ). this cosmopolitan location provides easy (...)

**Example 6.3 — Responses by ChatGPT (Description profile)****Response with low groundedness to description:**

“Thank you for your positive review! We’re glad you enjoyed your stay at Quest Bondi Junction . Our rooms are indeed well positioned, clean, and quiet , offering a peaceful environment for our guests. Our friendly staff takes pride in providing excellent service to ensure a pleasant experience . We appreciate your feedback regarding the TV remote , and we’re pleased to hear that it was immediately fixed by our attentive staff. The hotel’s location near the train station and public transport , as well as its proximity to shopping and food outlets , adds to the convenience for our guests. We also offer free Wi-Fi, with the option of higher speeds at an additional cost . We look forward to welcoming you back on your next trip.”

**Response with high groundedness to description:**

“Thank you for taking the time to share your positive experience at Quest Bondi Junction . As a provider of contemporary accommodation, we are committed to delivering exceptional service and ensuring guest satisfaction . Our 82 spacious studio apartments are thoughtfully designed , combining comfort and style to meet the needs of both corporate and leisure travelers . With our anticipated 4½ star rating , we aim to exceed expectations and create memorable stays. Our location within walking distance of Westfield Shopping Centre , Centennial Park , and major public transport networks added to the convenience of your stay. We appreciate your acknowledgment of our well-positioned, clean, and quiet rooms , which offer a serene environment for relaxation. Our dedicated staff takes pride in their friendliness and responsiveness, promptly addressing any concerns or requests, as evidenced by your TV remote issue . We are pleased to provide complimentary Wi-Fi , and for those requiring faster speeds, we offer options at an additional cost . Thank you for choosing Quest Bondi Junction, and we eagerly await the opportunity to welcome you back on your next trip, ensuring another exceptional experience.”

### 6.2.2.2 Amenities Profile Analysis

We next turn our focus to the Amenities profile, analyzing how both our KB100 and ChatGPT perform when generating responses grounded to amenities text. We delve into instances where our model shows minor improvements in informativeness using the Amenities profile. By examining these instances, we can discern response patterns of our model in Amenities profiles and identify effective strategies for further improvement.

Example 6.4 provides an instance of a review and amenities text.

#### Example 6.4 — An example of a review and amenities

**Review:**

“ramada inn cruise port —SEP— we have stayed here at least <DIGIT> times . the hotel is a great location for our cruises and great employees . we love the pool and the tiki bar . yes , it is an old hotel but every time we come back they are making improvements .”

**Amenities:**

free parking, free high speed internet ( wifi ), pool, fitness center with gym / workout room, free breakfast, billiards, free airport transportation, business center with internet access, parking, wifi, outdoor pool, bar / lounge, restaurant, breakfast buffet, vending machine, poolside bar, free shuttle or taxi services, airport transportation, shuttle bus service , taxi service, conference facilities, banquet room, meeting rooms, fax / photocopying, baggage storage

Example 6.5 presents responses generated by KB100, with the bias allocated to the amenities text in Example 6.4.

#### Example 6.5 — Responses by KB100 (Amenities profile)

**Response with Amenities 5:**

“<GREETING> thanks for the nice comments. i’m glad you had a nice stay with us here at the ramada airport / cruise port. we’re glad you liked our clean rooms, friendly staff, and shuttle service . thanks again for your nice comments and we hope to see you again soon. <SALUTATION>”

**Response with Amenities 50:**

“<GREETING> thanks for staying with us here at the ramada airport / cruise port. we’re glad you enjoyed your stay and thanks for the nice comments about our hotel, shuttle service , pool , and shuttle service . we hope you ’ll stay with us again the next time you’re in town. <SALUTATION>”

The responses exhibit some degree of hallucination while trying to incorporate certain information in the review and the amenities. For instance, the response with Amenities 5 mentions that “we’re glad you liked our (...) shuttle service”. Here,



*shuttle service* can be seen as informative tokens taken from the amenities, yet, exhibiting an incorrect fact in the context, as the reviewer did not mention this.

This type of hallucination was observed across the responses generated by our models with Amenities profiles. Moreover, as revealed in the results of our experiment, the responses from Amenities 5 and Amenities 50 do not show discernible differences in terms of the incorporation of amenities.

Example 6.6 displays responses generated by ChatGPT, directed to ground on the amenities above.

### Example 6.6 — Responses by ChatGPT

#### Response with low groundedness to amenities:

“Thank you for your review! We’re glad to hear that you enjoyed your stay at Ramada Inn Cruise Port. Our employees take pride in providing great service to our guests. **The pool and tiki bar** are popular amenities among our visitors. We appreciate your understanding that **the hotel is older**, but we’re constantly working on making improvements to enhance your experience.”

#### Response with high groundedness to amenities:

“Thank you for **being a loyal guest** and choosing Ramada Inn Cruise Port for your cruise vacations. We truly value your continued patronage. Our prime location near the cruise port makes it convenient for embarkation and disembarkation. Our dedicated team members take pride in providing excellent service and creating a welcoming atmosphere for our guests. We’re thrilled to hear that you enjoyed **the pool and tiki bar**, which are popular amenities to relax and socialize. We acknowledge that the hotel has **a rich history**, and we constantly invest in renovations and improvements to ensure a comfortable and enjoyable stay for our valued guests like you.”

These responses, similar to those in the Description profile, address key aspects from the review, such as the *pool*, the *tiki bar*, that the hotel is *old* and that the reviewer is a *loyal guest*. However, in this instance, the model failed to incorporate extra amenities properly into the responses. Moreover, they exhibit a degree of hallucination as well, e.g. “*Our prime location near the cruise port makes it convenient for embarkation and disembarkation*”, which is not mentioned in either the review or amenities.

This pattern was noticeable across several samples generated by ChatGPT. The responses either neglected to incorporate amenities noticeably as seen in this example, or merely listed them without integrating smoothly into the response (identified by phrases like “*Our amenities include ...*”, “*You can take advantage of our amenities such as ...*”). In conclusion, compared to grounding on description, the ChatGPT model had more difficulty generating pertinent responses that integrate required attributes without hallucination when grounded on amenities.

### 6.2.2.3 Knowledge Profile Analysis

Finally, we turn our attention to the Knowledge profile, where the ChatGPT model is prompted to blend information from both the description and amenities. For this analysis, we present the same example as in the Description profile analysis. The example of a review, description and amenities are provided in Example 6.7.

#### Example 6.7 — An example of a review, description, and amenities

##### Review:

“well positioned , nice clean and quiet room, friendly staff , good price , what more can you ask ? —SEP— very happy with the room and features , small problem with the tv remote immediately fixed upon drawing to the attention of staff , how would they know otherwise . the quest bondi junction is conveniently located in close proximity to the train station and public transport with lot of shopping and food outlets nearby . free wifi , higher speed available at additional cost . would definitely stay again on the next trip down .”

##### Description:

quest bondi junction provides contemporary accommodation that delivers the excellent standards of service and guest experiences you have come to enjoy and expect from quest serviced apartments. the 82 spacious studio apartments are well equipped and stylishly furnished to satisfy the needs of the corporate and leisure traveller. quest bondi junction has an anticipated 4½ star rating and is conveniently located within walking distance of westfield shopping centre, centennial park and major public transport networks ( rail and bus ). this cosmopolitan location provides easy (...)

##### Amenities:

paid private parking on - site, free high speed internet ( wifi ), wifi, breakfast available, taxi service, non - smoking hotel, laundry service, self - serve laundry, air conditioning, desk, housekeeping, seating area, kitchenette, dishwasher, cable / satellite tv, walk - in shower, sofa, telephone, wake - up service / alarm clock, microwave, refrigerator, electric kettle, flatscreen tv, radio, complimentary toiletries, hair dryer,

#### Example 6.8 — Responses by KB100 (Knowledge profile)

##### Response with Knowledge 10:

“<GREETING> thank you for taking the time to review your recent stay with us at quest bondi junction . we value all of our guest feedback as it allows us the opportunity to reflect on our services and facilities. i am very pleased to hear that you enjoyed your stay and the convenience of our location . we look forward to welcoming you back to <LOC> in the near future. <SALUTATION>”

##### Response with Knowledge 50:

“<GREETING> thank you for taking the time to review your recent stay with us at quest bondi junction . we appreciate all of our guest feedback as it allows us the opportunity to reflect on our services and facilities. as you have mentioned, <LOC> junction is conveniently located in the heart of <NAME> with easy access to public transport , centennial park and the sydney transport network . i look forward to welcoming you back to the hotel in the near future. <SALUTATION>”

We first examine the responses generated by our model under the Knowledge profile, depicted in Example 6.8. The response with Knowledge 10 shows little discernible use of information from the knowledge snippets, with the exception of the hotel’s name taken from the description. The response references the hotel’s *convenient location*, a detail mentioned in both the review and the description, yet the remainder of the responses remains largely generic. This is distinct from the response with Description 10 (in Example 5.2), which exhibits a reasonable degree of grounding in the description.

The response with Knowledge 50 does show some grounding in the knowledge snippet, incorporating aspects such as *centennial part* and *transport networks*, but the majority of the grounding comes from the description rather than the amenities.

This tendency is a common trait across responses with Knowledge profiles from our models. Generally, they exhibit weak incorporation of knowledge. However, when knowledge is incorporated, it primarily originates from the description, indicating that a balance between both knowledge sources is yet to be achieved.

Example 6.9 shows the responses generated by ChatGPT with a Knowledge profile.

### Example 6.9 — Responses by ChatGPT (Knowledge profile)

#### Response with low groundedness to knowledge:

“Thank you for your positive review! We’re glad you enjoyed your stay at Quest Bondi Junction . Our staff takes great pride in providing friendly service and ensuring a comfortable experience for our guests. We apologize for the small issue with the TV remote but are pleased to hear that it was promptly resolved. Our location near the train station and public transport , as well as the nearby shopping and food outlets , makes us a convenient choice for travelers. We offer free Wi-Fi, and higher speed options are available for an additional cost . We look forward to welcoming you back on your next trip.”

#### Response with high groundedness to knowledge:

“Thank you for taking the time to share your experience at Quest Bondi Junction . We are delighted to hear that our well-positioned and clean rooms , along with our friendly staff , contributed to a memorable stay. Our commitment to excellent service is a priority, and we are pleased to have promptly resolved the issue with the TV remote as soon as it was brought to our attention. The convenience of our location near the train station, public transport , and various shopping and dining options is something we take pride in. We offer complimentary Wi-Fi to all our guests, and for those looking for higher speeds, we have an option available at an additional cost. We truly appreciate your kind words and look forward to welcoming you back on your next visit to Quest Bondi Junction.”

Responses with the Knowledge profile from ChatGPT exhibited a similar trend to those with the Description profile, addressing issues raised in the review, such as *TV remote*, *location near the public transport*, *shopping and food outlets*, and *Wi-Fi*.

Notably, as the groundedness of the response increases, so does the length of the response, yet the grounding leans more heavily towards the review. For instance, the high groundedness response brings up *well-positioned and clean rooms* from the review as additional detail. This pattern was noticed across responses under the Knowledge profile from ChatGPT, validating the decrease in precision and increase in recall observed in this profile.

### 6.2.3 Implications

Our automatic evaluation metrics clearly indicate that ChatGPT consistently outperforms our model in terms of informativeness. Furthermore, the qualitative analysis discovered another strength of ChatGPT to aptly incorporate aspects from the reviews into its generated responses. This additional incorporation significantly enhances the relevance and applicability of its responses.

Nonetheless, both models exhibit a degree of hallucination that introduces information not grounded in the provided inputs. Furthermore, when tasked with grounding on amenities, both models faced difficulties in integrating information from amenities properly. The transformer-based architectures such as ChatGPT and our model appear to struggle with handling inputs like amenities that lack contextual information.

These observations reveal a common challenge in knowledge-enhanced review response generation. Maintaining strict factual accuracy, while simultaneously promoting informativeness in the responses remains a demanding task, underscoring an area that needs further exploration and improvement.

## 6.3 Limitations and Future Directions

Our study marks a step towards advancing knowledge-enhanced review response generation using attention biasing mechanisms. However, there are limitations inherent in our research. These will provide opportunities for future exploration and development.

**Need for Human Evaluation:** While our study extensively utilized various automatic metrics to evaluate the models' performance, these automated metrics are not entirely reliable and might not capture the full spectrum of qualitative nuances existing in the generated responses. Factors such as coherence, relevance, and fluency, which are hard to measure using automatic metrics, might have been overlooked.

Future directions could involve augmenting our existing evaluation with human evaluation. This would provide more comprehensive overview of the model performances and help in validating the automated metrics based on human judgment.

**Interpretation of Perplexity:** In contrast to the preliminary experiment, our interpretation of increased perplexity - when accompanied by enhanced informativeness and diversity - has shifted. We considered it as a sign of less generic responses rather than a decrease in fluency. However, it is important to acknowledge the limitations of this interpretation. Higher perplexity signifies that the model’s output is less predictable. Although we presumed that this less predictable nature emanates from the model generating more interesting and less generic responses, we cannot dismiss the possibility that it may also indicate a decline in text quality. In other words, higher perplexity in our results could potentially imply reduced fluency, as poorer quality text would also be challenging for the model to predict. Therefore, a comprehensive exploration of perplexity’s implications is warranted to gain a deeper understanding of its relationship with both fluency and specificity.

**Truncation of Knowledge Snippets:** The KB settings in our study were fed with truncated knowledge snippets. This does not consider sentence-level linguistic boundaries. As a consequence, the majority of the knowledge text fed in these models are cut unnaturally<sup>14</sup>. This truncation method may have negatively affected the quality of input and the subsequent performance of the models. Future work could refine this truncation approach, selecting the most contextually rich and relevant parts of the text while preserving linguistic boundaries. We anticipate that such an approach would enhance model performance.

**Restricted Model and Language Scope:** Our study focused exclusively on the BART model and the English dataset, which may limit the generalizability of our findings. Future work could extend the scope by utilizing other Transformer-based models such as GPT-3 or T5, and exploring other languages to confirm the universality of our findings.

**Limited Analysis of Knowledge Input Length:** While our study examined the implications of changing knowledge input lengths, we did not conduct a comprehensive analysis of how these lengths might affect the model’s performance. Future work could undertake a more detailed investigation to determine the optimal length that would most effectively boost informativeness. By combining this with techniques

---

<sup>14</sup>The knowledge snippets in the OD and FD models are also truncated, as they use a fixed bucket size as well. However, the bucket sizes in these models correspond to the 75th quantile of the token length distribution in the dataset. Thus, the majority of the knowledge snippets are fed in full length.

to retrieve the most pertinent knowledge snippet of an appropriate length, we posit that the effectiveness of attention biasing could be maximized.

**Manual Adjustment of Biasing Value:** The attention biasing method introduced in the original paper manually tunes the biasing value, which, although simple, might not yield optimal results. The most effective bias values for maximizing informativeness without compromising other factors such as fluency or relevance can vary based on the knowledge snippet lengths and task attributes. Hence, future research could consider developing a mechanism to dynamically or adaptively adjust the biasing value.

Addressing these limitations, especially dynamically adjusting biasing value, may necessitate additional control-specific training. This can be computationally expensive and may compromise the flexibility of the approach. Considering that the attention biasing method presented in this study is a zero-shot approach, not requiring any control-specific training, it serves as a viable, easily applicable strategy for enhancing informativeness in review response generation.

## 7 Conclusion

This thesis addressed the challenge of generating more informative responses to hotel reviews by leveraging the cross-attention biasing method presented by Hazarika et al. [2021]. This method modifies the attention distribution of an encoder-decoder model, reallocating attention to different parts of the input.

We examined this attention biasing approach using the BART model under a variety of experimental conditions. We fine-tuned the model on a dataset consisting of hotel review response pairs, integrating two parts of knowledge as additional input. During the inference stage, we adjusted the model’s cross-attention weights towards these additional inputs, aiming to produce more informative responses grounded to the knowledge texts.

Our initial experiment aimed to reproduce the results of the original study by Hazarika et al. [2021], targeting a knowledge-enhanced dialog generation task. We successfully reproduced the results, reinforcing our confidence in the proposed methodology and forming a solid foundation for the subsequent experiments.

The foremost objective of our research was to determine the extent to which the cross-attention biasing could be applied to improve the informativeness of review responses in the hotel domain (RQ1). Our first two primary experiments indicated that, contrary to our initial expectations and the results from our preliminary experiment, the biasing technique failed to enhance the informativeness in our knowledge-enhanced review response generation task, even when fine-tuned with a filtered dataset exhibiting less genericness.

This discrepancy led us to our second research question (RQ2), where we explored the influence of the knowledge input length during the inference stage on the model’s performance with respect to informativeness. Our findings indicated that the cross-attention biasing method significantly enhances and controls the informativeness of the generated responses in settings where the input length of the provided knowledge snippet is restricted and the bias is allocated to the knowledge text that offers contextual and narrative information. Consequently, this affirms RQ1, but with the added caveat that the efficacy of the biasing method is largely reliant on specific

configurations.

We also performed a comparative analysis with ChatGPT. While ChatGPT outperformed our model in terms of informativeness, it was observed that both models showed a degree of hallucination and struggled to incorporate knowledge snippets that lack contextual information. This suggests the direction for further research and improvement.

In conclusion, this thesis represents a step towards understanding how to effectively apply the attention biasing knob to enhance informativeness in knowledge-enhanced review response generation. Although there remain open questions, such as determining an optimal bias value, integrating discrete information, and handling extended knowledge, the insights gained through our exploration are encouraging, shedding light on the potential future developments in the domain of knowledge-enhanced generation.



# 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, 2018.
- S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- A. Belz and E. Reiter. Comparing automatic and human evaluation of nlg systems. In *11th conference of the european chapter of the association for computational linguistics*, pages 313–320, 2006.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.
- A. S. Cantalops and F. Salvi. New consumer behavior: A review of research on ewom and hotels. *International Journal of Hospitality Management*, 36:41–51, 2014.
- D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.
- J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354, 2006.
- K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- J. Cohen. *Statistical power analysis for the behavioral sciences*. 1988.

- M. De Bruyn, E. Lotfi, J. Buhmann, and W. Daelemans. Bart for knowledge grounded conversations. *Converse@ KDD*, 2666, 2020.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.
- C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38, 2022.
- 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. IEEE, 2019.
- C. Gao, W. Zhou, X. Xia, D. Lo, Q. Xie, and M. R. Lyu. Automating app review response generation based on contextual knowledge. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(1):1–36, 2021.
- M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, and D. Hakkani-Tür. Topical-chat: Towards knowledge-grounded open-domain conversations. 2019.
- S. Hassan, C. Tantithamthavorn, C.-P. Bezemer, and A. E. Hassan. Studying the dialogue between users and developers of free apps in the google play store. *Empirical Software Engineering*, 23:1275–1312, 2018.
- D. Hazarika, M. Namazifar, and D. Hakkani-Tür. Zero-shot controlled generation with encoder-decoder transformers. *arXiv preprint arXiv:2106.06411*, 2021.
- B. Hedayatnia, K. Gopalakrishnan, S. Kim, Y. Liu, M. Eric, and D. Hakkani-Tur. Policy-driven neural response generation for knowledge-grounded dialogue systems. *arXiv preprint arXiv:2005.12529*, 2020.
- A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

- Z. Ji, Z. Lu, and H. Li. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*, 2014.
- 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.
- D. Katsiuba, T. Kew, M. Dolata, and G. Schwabe. Supporting online customer feedback management with automatic review response generation. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, 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 (ECNLP 5)*, pages 121–133, 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, 2020.
- 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. *arXiv preprint arXiv:1910.13461*, 2019.
- C. Li, G. Cui, and L. Peng. The signaling effect of management response in engaging customers: A study of the hotel industry. *Tourism Management*, 62: 42–53, 2017.
- C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*, 2023.
- R. Mir, B. Felbo, N. Obradovich, and I. Rahwan. Evaluating style transfer for text. *arXiv preprint arXiv:1904.02295*, 2019.
- 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, 2002.
- 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, Sept. 2015. Association for Computational

- Linguistics. doi: 10.18653/v1/W15-3049. URL <https://aclanthology.org/W15-3049>.
- M. Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4770. URL <https://aclanthology.org/W17-4770>.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- H. Rashkin, D. Reitter, G. S. Tomar, and D. Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. *arXiv preprint arXiv:2107.06963*, 2021.
- V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC<sup>2</sup> Workshop*, 2019.
- P. Shaw, J. Uszkoreit, and A. Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.
- S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015.
- I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.
- A. A. Syed, F. L. Gaol, and T. Matsuo. A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, 9:13248–13265, 2021.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Z. Wu, M. Galley, C. Brockett, Y. Zhang, X. Gao, C. Quirk, R. Koncel-Kedziorski, J. Gao, H. Hajishirzi, M. Ostendorf, et al. A controllable model of grounded response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14085–14093, 2021.
- B. Yang, Z. Tu, D. F. Wong, F. Meng, L. S. Chao, and T. Zhang. Modeling localness for self-attention networks. *arXiv preprint arXiv:1810.10182*, 2018.

- S. Yang, Y. Wang, and X. Chu. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*, 2020.
- W. You, S. Sun, and M. Iyyer. Hard-coded gaussian attention for neural machine translation. *arXiv preprint arXiv:2005.00742*, 2020.
- W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, and M. Jiang. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38, 2022.
- H. Zhang, H. Song, S. Li, M. Zhou, and D. Song. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*, 2022.
- T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- 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*, pages 2425–2435, 2019.
- Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu. Taxygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.

# A Tables

Standard deviation for OD in Experiment 1

	<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
	<b>PPL</b>	<b>Self-BLEU</b> ↓	<b>ChrF</b> <sub><i>t</i><sub>g</sub></sub>	<b>ChrF</b> <sub><i>k</i></sub>	<b>BERT</b> <sub><i>k</i></sub>	<b>Precision</b> <sub><i>k</i></sub>	<b>Recall</b> <sub><i>k</i></sub>
<b>Baseline</b>	27.671	0.060	0.049	0.012	0.003	0.014	0.012
<b>Review 5</b>	26.723	0.112	0.057	0.011	0.003	0.029	0.012
<b>Amenities 2</b>	27.573	0.080	0.043	0.017	0.006	0.031	0.017
<b>Amenities 5</b>	26.763	0.054	0.055	0.016	0.006	0.030	0.014
<b>Amenities 10</b>	25.356	0.136	0.034	0.011	0.002	0.027	0.006
<b>Amenities 50</b>	21.717	0.053	0.030	0.007	0.004	0.034	0.006
<b>Description 2</b>	28.404	0.092	0.055	0.010	0.004	0.048	0.008
<b>Description 5</b>	28.555	0.061	0.059	0.013	0.003	0.023	0.014
<b>Description 10</b>	27.963	0.063	0.039	0.016	0.005	0.042	0.009
<b>Description 50</b>	26.996	0.094	0.036	0.005	0.005	0.038	0.008
<b>Knowledge 2</b>	22.442	0.066	0.060	0.012	0.005	0.024	0.011
<b>Knowledge 5</b>	22.759	0.080	0.047	0.010	0.003	0.033	0.008
<b>Knowledge 10</b>	27.462	0.046	0.011	0.006	0.005	0.024	0.010
<b>Knowledge 50</b>	27.691	0.141	0.013	0.024	0.005	0.034	0.016

Table 10: The standard deviation for OD in Experiment 1

Standard deviation for FD in Experiment 2

	<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
	<b>PPL</b>	<b>Self-BLEU</b> ↓	<b>ChrF</b> <sub><i>t<sub>g</sub></i></sub>	<b>ChrF</b> <sub><i>k</i></sub>	<b>BERT</b> <sub><i>k</i></sub>	<b>Precision</b> <sub><i>k</i></sub>	<b>Recall</b> <sub><i>k</i></sub>
<b>Baseline</b>	28.118	0.039	0.080	0.021	0.006	0.020	0.020
<b>Review 5</b>	28.065	0.050	0.048	0.019	0.004	0.012	0.012
<b>Amenities 2</b>	27.648	0.058	0.036	0.009	0.003	0.025	0.008
<b>Amenities 5</b>	27.171	0.054	0.022	0.010	0.004	0.010	0.009
<b>Amenities 10</b>	26.011	0.039	0.041	0.008	0.002	0.040	0.007
<b>Amenities 50</b>	21.631	0.070	0.026	0.008	0.003	0.024	0.008
<b>Description 2</b>	28.569	0.071	0.041	0.013	0.002	0.038	0.015
<b>Description 5</b>	28.861	0.055	0.073	0.023	0.005	0.043	0.021
<b>Description 10</b>	28.663	0.021	0.050	0.010	0.003	0.020	0.005
<b>Description 50</b>	26.593	0.080	0.043	0.015	0.005	0.033	0.015
<b>Knowledge 2</b>	28.593	0.069	0.045	0.017	0.004	0.022	0.016
<b>Knowledge 5</b>	23.544	0.080	0.044	0.009	0.004	0.042	0.011
<b>Knowledge 10</b>	28.246	0.106	0.027	0.008	0.003	0.025	0.011
<b>Knowledge 50</b>	28.175	0.058	0.038	0.012	0.005	0.023	0.007

Table 11: The standard deviation for FD in Experiment 2

Results for Experiment 3 (KB40)

	<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
	<b>PPL</b>	<b>Self-BLEU</b> ↓	<b>ChrF<sub>tg</sub></b>	<b>ChrF<sub>k</sub></b>	<b>BERT<sub>k</sub></b>	<b>Precision<sub>k</sub></b>	<b>Recall<sub>k</sub></b>
<b>Baseline</b>	63.4 (0.363)	87.0 (0.082)	28.0 (0.042)	4.8 (0.015)	77.2 (0.003)	28.3 (0.028)	4 (0.009)
<b>Amenities 2</b>	64.6 (0.370)	86.9 (0.097)	27.2 (0.043)	4.6 (0.013)	77.2 (0.006)	28.2 (0.027)	3.8 (0.005)
<b>Amenities 5</b>	66.8 (0.250)	87.4 (0.046)	26.0 (0.042)	4.3 (0.010)	77.1 (0.006)	28.1 (0.040)	3.5 (0.009)
<b>Amenities 10</b>	68.3 (0.276)	87.8 (0.061)	25.0 (0.042)	4.0 (0.010)	77 (0.006)	27.9 (0.040)	3.3 (0.009)
<b>Amenities 50</b>	72.0 (3.337)	88.8 (0.121)	23.4 (0.034)	3.7 (0.005)	76.9 (0.003)	27.6 (0.053)	3.1 (0.008)
<b>Description 2</b>	67.6 (0.426)	86.5 (0.165)	28.2 (0.030)	4.9 (0.008)	77.2 (0.002)	28.7 (0.030)	4.1 (0.007)
<b>Description 5</b>	73.9 (0.248)	85.8 (0.031)	28.3 (0.028)	5.1 (0.004)	77.3 (0.004)	29.6 (0.035)	4.2 (0.006)
<b>Description 10</b>	80.1 (0.353)	85.2 (0.084)	28.1 (0.013)	5.2 (0.007)	77.4 (0.005)	30.3 (0.046)	4.4 (0.008)
<b>Description 50</b>	99.3 (0.500)	82.0 (0.059)	27.3 (0.039)	5.3 (0.004)	77.5 (0.004)	32.2 (0.030)	4.7 (0.005)
<b>Knowledge 2</b>	68.9 (0.303)	86.5 (0.113)	27.5 (0.048)	4.7 (0.012)	77.2 (0.006)	28.6 (0.049)	3.9 (0.009)
<b>Knowledge 5</b>	78.6 (0.602)	86.2 (0.090)	26.8 (0.045)	4.6 (0.007)	77.2 (0.005)	29.1 (0.052)	3.8 (0.009)
<b>Knowledge 10</b>	86.0 (0.401)	85.7 (0.146)	26.3 (0.056)	4.5 (0.013)	77.2 (0.005)	29.5 (0.046)	3.8 (0.011)
<b>Knowledge 50</b>	100.1 (0.786)	84.3 (0.149)	25.9 (0.019)	4.6 (0.019)	77.3 (0.010)	30.2 (0.029)	3.9 (0.010)

Table 12: Evaluation metrics for the KB40 in Experiment 3. ChrF refers to ChrF++ and BERTScore denotes the F1 score. The reported results are the average across five random seeds. Standard deviations are reported in parenthesis.



Standard deviation for Experiment 3 (KB100)

	<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
	<b>PPL</b>	<b>Self-BLEU</b> ↓	<b>ChrF<sub>tg</sub></b>	<b>ChrF<sub>k</sub></b>	<b>BERT<sub>k</sub></b>	<b>Precision<sub>k</sub></b>	<b>Recall<sub>k</sub></b>
<b>Baseline</b>	0.404	0.080	0.031	0.012	0.003	0.023	0.015
<b>Amenities 2</b>	0.317	0.068	0.041	0.021	0.004	0.043	0.010
<b>Amenities 5</b>	0.237	0.116	0.032	0.006	0.002	0.013	0.005
<b>Amenities 10</b>	0.320	0.093	0.029	0.006	0.005	0.038	0.006
<b>Amenities 50</b>	0.222	0.079	0.035	0.018	0.005	0.023	0.012
<b>Description 2</b>	0.196	0.102	0.050	0.022	0.005	0.019	0.017
<b>Description 5</b>	0.417	0.067	0.047	0.005	0.003	0.036	0.005
<b>Description 10</b>	0.418	0.060	0.030	0.008	0.004	0.027	0.003
<b>Description 50</b>	0.324	0.059	0.019	0.009	0.003	0.027	0.014
<b>Knowledge 2</b>	0.449	0.123	0.040	0.010	0.004	0.031	0.010
<b>Knowledge 5</b>	0.212	0.042	0.023	0.010	0.003	0.028	0.008
<b>Knowledge 10</b>	0.430	0.100	0.034	0.011	0.004	0.055	0.009
<b>Knowledge 50</b>	0.589	0.069	0.036	0.009	0.006	0.040	0.007

Table 13: The standard deviation for Experiment 3 (KB100)

Results for Experiment 3 (KB160)

	<i>Perplexity</i>	<i>Diversity</i>	<i>Gold Alignment</i>	<i>Informativeness</i>			
	<b>PPL</b>	<b>Self-BLEU</b> ↓	<b>ChrF</b> <sub>lg</sub>	<b>ChrF</b> <sub>k</sub>	<b>BERT</b> <sub>k</sub>	<b>Precision</b> <sub>k</sub>	<b>Recall</b> <sub>k</sub>
<b>Baseline</b>	84.6 (0.384)	77.6 (0.081)	33.9 (0.038)	6.1 (0.023)	77.3 (0.004)	28.5 (0.031)	4.7 (0.019)
<b>Amenities 2</b>	85.8 (0.446)	77.3 (0.127)	33.3 (0.073)	5.9 (0.022)	77.3 (0.004)	28.6 (0.021)	4.6 (0.018)
<b>Amenities 5</b>	87.8 (0.241)	77.7 (0.101)	32.0 (0.023)	5.5 (0.005)	77.3 (0.002)	28.6 (0.018)	4.3 (0.005)
<b>Amenities 10</b>	87.2 (0.557)	78.7 (0.064)	30.9 (0.031)	5.2 (0.031)	77.2 (0.006)	28.7 (0.035)	4.1 (0.008)
<b>Amenities 50</b>	80.5 (0.133)	81.2 (0.027)	28.2 (0.021)	4.6 (0.010)	77.2 (0.003)	28.7 (0.036)	3.7 (0.009)
<b>Description 2</b>	86.7 (0.385)	77.6 (0.037)	34.0 (0.061)	6.2 (0.023)	77.4 (0.006)	28.8 (0.021)	4.8 (0.016)
<b>Description 5</b>	88.6 (0.665)	77.9 (0.079)	33.9 (0.030)	6.3 (0.015)	77.4 (0.006)	29.4 (0.037)	5 (0.012)
<b>Description 10</b>	89.2 (0.555)	78.2 (0.030)	33.8 (0.063)	6.4 (0.011)	77.5 (0.002)	29.8 (0.041)	5.1 (0.008)
<b>Description 50</b>	89.3 (0.512)	79.5 (0.105)	32.8 (0.040)	6.4 (0.008)	77.6 (0.004)	30.7 (0.026)	5.2 (0.007)
<b>Knowledge 2</b>	87.7 (0.489)	77.2 (0.109)	33.5 (0.032)	6.0 (0.009)	77.4 (0.001)	28.8 (0.012)	4.7 (0.003)
<b>Knowledge 5</b>	92.2 (0.381)	77.0 (0.042)	32.9 (0.054)	5.8 (0.007)	77.4 (0.002)	29.1 (0.007)	4.6 (0.009)
<b>Knowledge 10</b>	94.5 (0.214)	77.0 (0.106)	32.3 (0.063)	5.8 (0.011)	77.4 (0.003)	29.3 (0.022)	4.5 (0.006)
<b>Knowledge 50</b>	97.7 (0.290)	77.4 (0.083)	31.6 (0.027)	5.6 (0.004)	77.4 (0.001)	29.7 (0.025)	4.5 (0.008)

Table 14: Evaluation metrics for the KB160 in Experiment 3. ChrF refers to ChrF++ and BERTScore denotes the F1 score. The reported results are the average across five random seeds. Standard deviations are reported in parenthesis.