AI-Seminar

TECHNISCHE UNIVERSITÄT DARMSTADT

Enhancing efficiency in job matching apps through AI-suggested chat answers based on data from past conversations

submitted at

Department of Law and Economics

Chair of Marketing & Human Resource Management

Univ.-Prof. Dr. Dr. Ruth Stock-Homburg

Supervisor:

Nils Schönfeld

Technical University of Darmstadt

Winter term 2023/24

written by

Kristian Efa

Matriculation number: 2519036

Jan Zimny

Matriculation number: 2921556

Abstract

This seminar thesis focuses on the development and evaluation of a model designed for generating contextually enriched smart replies within the job matching domain, utilizing context from previous chats. Through a series of experiments, participants assumed specific personas and employed the developed smart reply model to respond to context-dependent chats. Participants had the freedom to edit the smart replies as needed, and their alterations were tracked. The research concludes with a comprehensive qualitative and quantitative analysis of the editing behavior and participant evaluations of the model. The findings of the thesis confirm that the outlined model demonstrates proficiency in contextualizing smart replies for a high conversational relevance in a job matching environment. Furthermore, the study finds that participants expressed satisfaction, perceived the smart replies to be natural, and illustrated a high likelihood of using such. This research contributes valuable insights for enhancing communication in job matching scenarios and lays the groundwork for future advancements in the field.

Inhaltsverzeichnis

Table of Figures

Table of Tables

Abkürzungsverzeichnis

AI	Artificial Intelligence
AI-MC	AI-Mediated Communication
GPT	Generative Pre-Trained Transformer
H	Hypothesis
LLM	Large Language Model
LSR	LLM-generated Smart Replies
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
RQ	Research Question

1 Introduction

1.1 Practical Relevance

In the recent years, natural language processing (NLP) has become a focal point, propelled by remarkable advancements in AI models (Koubaa et al., 2023, p. 118698). According to PWC's global artificial intelligence (AI) study, the application of AI technology holds the potential to add an estimated $15.7 trillion to the global economy by the year 2030 (PWC, 2017, p. 3). In addition to high economic value, the introduction of new AI technologies has a potentially big "... impact on knowledge work—especially activities involving decision making and collaboration. …. This is because of [their] ability to predict patterns in natural language and use it dynamically" (McKinsey & Company, 2023). While early AI assistants encountered limitations due to limited knowledge bases and deterministic decision models, the development of modern large language models (LLMs) has led to evolving language and knowledge capabilities (Bastola et al., 2023). This is further underlined by the success of OpenAI's Chat Generative Pre-Trained Transformer (ChatGPT). The launch of ChatGPT in November 2022 caused a surge of public interest due to its tremendous ability to generate human-like sentences by leveraging deep learning techniques and massive data sets (OpenAI, 2022; Koubaa et al., 2023; Inoue et al., 2023). Language models such as ChatGPT are not limited to solely generating human-like sentences but can adapt to individual conversation styles and generate logically coherent and personalized replies, taking the conversation context into account (Bastola et al., 2023). By leveraging these capabilities and incorporating them into interpersonal text-based communication scenarios to generate smart replies, there is a significant potential to enhance overall communication productivity and efficiency (Bastola et al., 2023). Practical applications of AI-assisted response generation, for example, reported an increase in overall communication efficiency of 18.6% and a reduction of total response time of almost 50% (Bastola et al., 2017, p. 6; Fu et al., 2023, p. 5). The demand for response assistance underscores the necessity and appeal of such systems, even in less technologically advanced contexts. For example, despite generating shorter and less contextualized selections, Gmail's Smart Reply system was employed in 10 percent of all mobile email replies in 2016 (Kannan et al., 2016, p. 2). However, two months after its launch, OpenAI reported an impressive figure of over 100 million active monthly users, underscoring the sustained demand for personal text-based AI assistance (Reuters Media, 2023).

1.2 Scientific Relevance

The rapid evolution of LLMs and their integration into assisting interpersonal communication has prompted researchers to investigate their application in this domain. Prior studies employed different AI models but concentrated on evaluating the use of real-time response assistance in text-based conversations (Bastola et al., 2017; Buschek et al., 2021; Fu et al., 2023; Hancock et al., 2020; Hohenstein & Jung, 2018; Inoue et al., 2023; Mieczkowski et al., 2021). Several studies, for instance, analyzed absolute utilization frequencies, impact on communication efficiency, and performance in cognitive tasks (Bastola et al., 2017, p. 6; Fu et al., 2023, pp. 5–6; Hohenstein & Jung, 2018, p. 4; Inoue et al., 2023, p. 1975; Mieczkowski et al., 2021, p. 10). Individual perception of the assistance and users' editing behavior were additional points of interest (Fu et al., 2023, p. 6; Inoue et al., 2023, pp. 1976–1978). Studies that utilized the ChatGPT model for response assistance revealed positive impacts on characteristics like total response time, perceived helpfulness, and naturalness (Fu et al., 2023, pp. 5–7; Mieczkowski et al., 2021, p. 8). However, findings from previous studies consistently identify the absence of contextualization in generations as a primary drawback in using such assistance (Bastola et al., 2017, p. 7; Hohenstein & Jung, 2018, p. 6). A primary challenge in generating high-quality response selections is attributed to the limitation of input context, presenting a fundamental existing research gap (Fu et al., 2023, p. 10).

1.3 Aims and Structure

Building on top of prior research, this study aims to further enhance interpersonal communication efficiency by employing well-contextualized AI-generated smart reply suggestions. The contextual scope of this research involves a job matching smartphone application developed by the Marketing department of the Technical University of Darmstadt (leap in time, 2023). Within this application, job applicants and employers are able to engage through a chat interface. The main objective of this study is to evaluate the use and perception of contextualized smart reply suggestions generated by a contemporary LLM within the context of job placement conversations. In terms of the scientific methodology, an unstructured literature analysis is undertaken as the initial step, laying the theoretical foundation for developing hypotheses and research questions in this study. The evaluation will take place in an experimental setting and will involve both qualitative and quantitative analyses. The experiment is facilitated through a basic chat interface integrated into a proposed technical implementation of a standalone software module. This module replicates chat conversations within the specified contextual scope, providing coherent reply suggestions to the user. The

remainder of this work is divided into distinct sections. The following section will define core concepts and terms related to AI-mediated communication (AI-MC) and outline the current state of the literature. The third section will delineate the proposed technical implementation employed in the experiment. Subsequently, the fourth section will encompass the experiment and its results, followed by a section addressing implications, limitations, and areas of future research. The last section wraps up the thesis with a conclusion.

2 Theoretical Foundation

The following section outlines the current state of research regarding the application of AI-generated smart replies and their influence on communication efficiency and individual perception. Initially, a conceptual framework is introduced to categorize diverse AI-based mediation tools, accompanied by the definition of central terms in the context of this work. Subsequently, an overview of findings from prior research is provided, and research questions and hypotheses that this study endeavors to answer and verify are outlined.

2.1 Outline AI-MC Framework and Central Terms

With the advancement of powerful AI models, a diverse range of tools for mediating interpersonal text-based communication have emerged. These tools can propose, augment, modify, or generate messages to achieve an intended outcome. Examples encompass smart replies, auto-completion systems, auto-responses, auto-correct, predictive text, and grammar correction (Hancock et al., 2020, p. 90). Hancock et al. (2020) introduced a conceptual definition of AI-mediated communication (AI-MC), a framework that has gained widespread acceptance among researchers exploring the various applications and implications of AI assistance in interpersonal communication scenarios (Goldenthal et al., 2021, p. 2; Inoue et al., 2023, p. 1974; Jakesch et al., 2019, p. 1; Mieczkowski et al., 2021, p. 2). Hancock et al. (2020, p. 90) defines AI-MC as "… mediated communication between people in which a computational agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication or interpersonal goals" (Hancock et al., 2020, p. 90). Generally, there are various AI-MC tools with different mediating functions such as grammar correction, auto-responses or smart replies. To enhance comparability between various AI-MC tools and ensure mutual understanding, Hancock et al. (2020, p. 90) additionally presented a framework for categorizing various AI-MC tools. This framework encompasses five dimensions and seeks to systematically categorize a range of different AI-MC tools. The five dimensions include the magnitude of AI intervention, media type, optimization goal,

autonomy, and role orientation (Hancock et al., 2020, p. 91). The AI-based generation of smart replies in interpersonal communication is one of the core elements of AI-MC (Inoue et al., 2023, p. 1974). As a result, AI-MC tools that offer AI-generated smart replies can be categorized based on the AI-MC classification dimensions as text-based tools operating on behalf of a sender, with a high magnitude of AI involvement, varying optimization goals, and low AI autonomy (Hancock et al., 2020, pp. 91–92). As evident in recent scientific research and practical advancements of AI-MC tools, implementations of AI-based smart reply tools exhibit differences, resulting in a variety of assistance options (Buschek et al., 2021, p. 2; Fu et al., 2023, p. 10; Goldenthal et al., 2021, p. 2). Therefore, a precise delineation of these various options is crucial within the context of this research. In general, AI-MC reply suggestion tools can be categorized based on their output length, including distinctions such as sentence-level or message-level reply suggestions (Buschek et al., 2021; Fu et al., 2023, p. 10). While sentence-level suggestions assist users in seamlessly continuing already written text, message-level reply suggestions offer a thorough and complete response message to a received message (Fu et al., 2023, pp. 3–4). Furthermore, single-word suggestion tools, also known as predictive text suggestion tools, aim to predict a concise set of words with the highest probability of being typed next (Arnold et al., 2020, p. 128). As this work aims to enhance communication efficiency through comprehensive and contextual smart replies, options with a short output length, like single-word or sentence-level suggestions, are excluded from the investigation. Several recent studies predominantly utilized the two terms "smart replies" or "response suggestions" to denote AI-generated reply suggestions (Bastola et al., 2017, p. 2; Fu et al., 2023, p. 1; Inoue et al., 2023, p. 1974; Mieczkowski et al., 2021, p. 1). However, none of these studies offers a clear definition of the used term. To maintain consistency with prior research, this study adheres to the term "smart reply" and refines its definition based on the AI-MC definition as "suggested reply messages in text-based human conversations, generated by a LLM on behalf of the communicators to accomplish communication or interpersonal goals." Hancock et al. (2020) refer to AI as “computational systems that involve algorithms, machine learning methods, natural language processing, and other techniques that operate on behalf of an individual to improve a communication outcome” (Hancock et al., 2020, p. 90). In the context of this work, AI models are specified as LLMs. Hence, the present sutdy utilizes the term “LLM-based smart reply (LSR)”, as introduced by Bastola et al. (2017, p. 1), as general term to refer to smart replies independent of a specific underlying LLM. Previous AI-MC studies that focused on LSR mainly employed the GPT model to generate their reply suggestions (Bastola et al., 2017, p. 4; Fu et al., 2023, p. 1; Inoue et al., 2023, p. 1975). OpenAI’s GPT models are text-generation

models that generate text outputs depending on the received inputs. These inputs are designated as “prompts” and determine the outputs of the model (OpenAI, 2024e).

2.2 Literature Review

Recent empirical studies investigated AI-MC smart reply tools, from diverse perspectives. This section aims to provide a summary of recent scientific investigations regarding the application of AI-generated smart replies in interpersonal communication scenarios. Previous research in this domain primarily concentrated on assessing the influence of AI-generated smart replies on factors including frequency of use, editing behavior, interpersonal perception, and task accuracy (Bastola et al., 2017; Buschek et al., 2021; Fu et al., 2023; Inoue et al., 2023; Mieczkowski et al., 2021). Hohenstein and Jung (2018) conducted a comparison of dialogue histories between participant pairs using AI-assisted and standard messaging applications. The objective was to monitor the utilization rates of LSR. They employed LSRs of Google’s messaging app Allo. Discontinued in 2019, Google Allo was a publicly accessible AI-assisted messaging app that provided suggested responses derived from parsing the conversation history (Google, 2023). The smart replies typically appeared in groups of three after sending or receiving a message (Hohenstein & Jung, 2018, p. 2). Results revealed that the AI assistant’s smart replies were utilized merely 6.24% of the time (Hohenstein & Jung, 2018, p. 4). The outcomes of the qualitative interviews validated that the low utilization rates stemmed from inconsistencies between the suggested smart reply options and the conversation task. Consequently, the content of Allo’s LSRs frequently lacked relevance to the context of the conversation (Hohenstein & Jung, 2018, p. 4). Mieczkowski et al. (2021) employed Google’s smart reply system to investigate how they are integrated into a text-based communication task. In contrast to the study by Hohenstein and Jung (2018), their research aimed to investigate the impact of LSRs on language production, interpersonal perception, and task performance rather than absolute utilization frequency (Mieczkowski et al., 2021, p. 2) The task required participants to collaboratively match abstract images with their corresponding numbers by communicating through the Google Hangouts Chat messenger. Regarding language production Mieczkowski et al. (2021) were specifically interested in the conversational conditions that led users to preferably use smart replies. Consequently, they analyzed the frequency of smart reply use across different conversational conditions, employing the adjacency pairs framework (Mieczkowski et al., 2021, p. 3; SACKS et al., 1974, p. 728). Adjacency pairs involve successive utterances by two individuals, where the second utterance is contingent upon the first one, such as a pair consisting of a question and an answer (Mieczkowski et al., 2021, p. 3).

The experiment results show that smart replies were predominantly employed within the context of question/answer adjacency pairs, accounting for 34% of the instances (Mieczkowski et al., 2021, p. 8). Despite being instructed to utilize appearing AI-generated smart replies, participants refused to use them 22.2% of the time, citing the lack of relevance of the provided suggestions to the given adjacency pair (Mieczkowski et al., 2021, p. 9). This aligns with the findings of the study conducted by Hohenstein and Jung (2018). However, in general, AI-generated smart replies were frequently integrated successfully as part of a response, leading to composite messages consisting of both human-generated and AI-generated text (Mieczkowski et al., 2021, p. 12). Regarding the impact on interpersonal perception, Mieczkowski et al. (2021) found that smart replies had no influence on the perception of warmth, competence, or task attraction for the receiver of messages. However, they did observe a decrease in the perception of social attraction (Mieczkowski et al., 2021, pp. 9–10). This is hypothesized to stem from the identified positivity bias of the previously provided smart replies (Mieczkowski et al., 2021, p. 12). Regarding the impact on task performance, they focused on examining task accuracy, length of conversation, and word count per message. Task accuracy was measured as a percentage share of correctly numbered tangrams per participant group, and conversation length reflected the total number of messages per group. The study findings revealed no significant influence of the use of smart replies regarding task accuracy and slight differences in conversational length. However, the utilization of AI-generated smart replies resulted in an average word count increase to 9.94 words per message, in contrast to conventionally written responses (Mieczkowski et al., 2021, p. 10). The outlined studies by Hohenstein and Jung (2018) and Mieczkowski et al. (2021) both employed Google's Smart Reply system, which is restricted to short suggestions with a low amount of incorporated context (Kannan et al., 2016). Despite Hohenstein and Jung (2018) identifying missing context as the primary reason for not using Google's smart replies, Mieczkowski et al. (2021, p. 8) demonstrated that humans are likely to integrate them into their own language under specific conversational conditions, such as question-answer pairs. Moreover, these limited response suggestions demonstrated a remarkable level of naturalness (Mieczkowski et al., 2021, p. 8). Advanced LLMs, such as the GPT model, enable longer and more effective contextualization, resulting in more complex suggestions (Bastola et al., 2017, p. 2). The language capabilities of the GPT-3 model, for instance, make it challenging for untrained human evaluators to accurately differentiate between GPT-3 generated outputs and human-generated texts (Brown et al., 2020, p. 9; Clark et al., 2021, p. 7285). Generally, the strength of LLMs resides in their ability to produce linguistically high-quality texts, while their weaknesses are primarily rooted in content-related

issues. These issues encompass the generation of false facts or inconsistencies concerning the context (Brown et al., 2020, p. 15; Clark et al., 2021, p. 7283). Clark et al. (2021, p. 7283) revealed that human evaluators primarily concentrate on surface-level aspects such as grammar, spelling, and style rather than focusing on the content in detail. Due to the sophisticated capabilities in generating fluent and linguistically high-quality texts, they therefore proposed a shift towards assessing the usefulness rather than "humanlikeness" when evaluating LSRs. The existing research landscape on the utilization of advanced LLMs for generating smart replies for communication assistance is relatively limited. Existing research has focused on modern LLMs, such as the GPT model, to generate contextualized smart replies, aiming to explore their utilization and impact on interpersonal communication (Bastola et al., 2017; Buschek et al., 2021; Fu et al., 2023; Inoue et al., 2023). Inoue et al. (2023) aimed to identify overall usage rates of smart replies generated by OpenAI's GPT-3 model. In the first part of their study, they consequently conducted an experiment in which participants were instructed to review pre-existing chat histories and then choose one of two presented response suggestions. One of the displayed suggestions had been generated by the GPT-3 model, while the other represented the authentic response found within the chat dialogue. Moreover, they were unable to discern the specific source of the two suggested smart replies. The prompt to generate the smart replies included the most recent four turns of the dialogue history from the chat corpus as context. The AI-generated smart replies were selected in 32.6% of instances. This finding indicates that GPT-3 generated response suggestions exhibit a certain degree of contextual consistency and naturalness (Inoue et al., 2023, p. 1975). The second part of the study by Inoue et al. (2023) aimed to investigate how GPT-3 generated response suggestions alter the dialogue flow in text-based interpersonal conversations. Therefore, dialogue acts such as question, answer, inform, directive, and commissive were employed. Dialogue acts represent the function of a written message in a dialogue (Inoue et al., 2023, p. 1976). To analyze how response suggestions alter the dialogue flow

- the influence of different dialogue acts on human editing tendency,
- and whether individuals modify the original dialogue act of smart replies when integrating them into their own replies was investigated.

The experimental set-up replicated that of the first experiment with a minor modification wherein participants were provided precisely one smart reply generated by GPT-3. They were instructed to edit the shown smart reply to create an appropriate response to the last message of the chat corpus (Inoue et al., 2023, p. 1976). The five dialogue acts question, answer, inform, directive, and commissive were equally distributed across all execution runs of the experiment.

To identify the editing behavior for each of these dialogue acts, Inoue et al. (2023) calculated BLEU scores for each of the 100 created responses (Inoue et al., 2023, pp. 1976–1977). The BLEU metric, initially developed for assessing machine translation quality, is an automated measure relying on modified n-gram precision calculations. It quantifies the extent to which an edited response incorporates identical words and expressions found in a suggested response. The BLEU score spans from 0 to 1, whereby a value close to 1 suggests almost no change, and a value close to 0 refers to a significant change to the response suggestion (Inoue et al., 2023, p. 1976; Papineni et al., 2002, pp. 3, 5; van der Lee et al., 2021, p. 3). Table 1 presents an overview of the average BLEU scores represented as percentages for dialogue acts examined by Inoue et al. (2023, p. 1977). As evident, smart replies that corresponded to the dialogue act 'answer' were edited the lowest on average according to the BLEU metric. This result aligns with the finding of Mieczkowski et al. (2021) that provided AI-generated smart replies are predominantly employed within the context of question/answer adjacency pairs (Mieczkowski et al., 2021, p. 8).

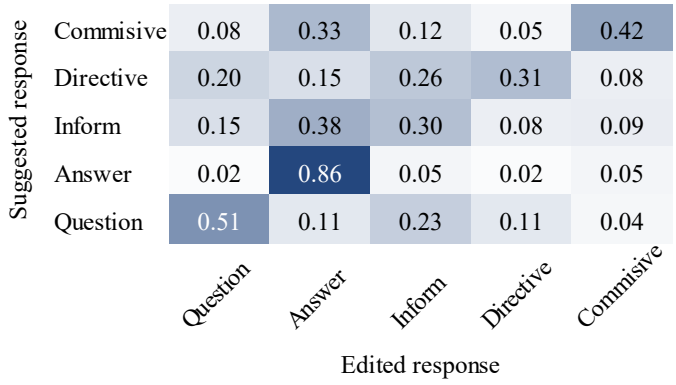

Figure 1: Frequency of transition between suggested responses and sent response for each pair of dialogue acts (Inoue et al., 2023, p. 1977)

Dialogue act	BLEU score
Question	21.76
Answer	61.92
Inform	18.75
Directive	28.86
Commissive	27.72

Table 1: Average BLEU scores per dialogue act (Inoue et al., 2023, p. 1977)

To assess whether participants altered the dialogue acts of the smart replies, Inoue et al. (2023) compared the dialogue act of the edited replies with the dialogue acts initially provided by the AI model. Their findings revealed that the original dialogue act was changed in 52% of the cases. This implies that people are willing to create diverse responses despite the presence of AI-generated smart replies (Inoue et al., 2023, p. 1977). In their third experiment, Inoue et al. (2023) aimed to investigate the importance of user initiative for the utilization of AI-generated smart replies. User initiative refers to the user's individual motivation participating in the dialogue to reach a dialogue goal. It was hypothesized that the frequency of utilization decreases when users take a more proactive role in the dialogue (Inoue et al., 2023, p. 1974). Following

that, participants engaged in two distinct chat sessions. During one session, they were instructed to collaboratively achieve the predetermined dialogue goal of crafting a fictional holiday plan together. Participants were provided with three AI-generated smart replies each time they had to respond. Afterward, average usage rates and BLEU scores were calculated per group and associated with insights derived from qualitative interviews (Inoue et al., 2023, pp. 1977–1978). On one hand, the findings reveal that content similarity between a suggestion and a user's response idea facilitates the utilization of smart replies. In these cases, smart replies act as an aid towards reaching that goal. Inoue et al. (2023) observed that participants whose response ideas aligned with the content of generated smart replies sometimes constructed composite messages by integrating portions of the suggested response into their own language to reduce input effort (Inoue et al., 2023, p. 1978). This observation aligns with findings from prior studies (Mieczkowski et al., 2021, p. 12). On the other hand, even if participants encountered uncertainty regarding the progression of the dialogue, smart replies were used as a source of inspiration (Inoue et al., 2023, pp. 1978–1979). Inoue et al. (2023) concluded that individuals with higher conversation motivation tend to avoid using response suggestions (Inoue et al., 2023, p. 1978). However, it's crucial to note limitations in their final experiment, including a small participant pool of only six males. Additionally, the study was conducted exclusively in Japanese, leading Inoue et al. (2023, p. 1979) to hypothesize potential variations in utilization behavior across different languages. Bastola et al. (2017) also utilized the GPT-3 model to generate smart replies. Similar to the study conducted by Mieczkowski et al. (2021), the objective was to analyze their impact on enhancing task accuracy. In the study by Mieczkowski et al. (2021), smart replies were incorporated into interpersonal text-based conversations directly associated with the collaborative task. Conversely, in the present outlined study, it was hypothesized that utilizing AI-generated smart replies in recurring formal conversations might reduce human engagement, consequently freeing up mental resources for non-communication tasks (Bastola et al., 2017, p. 2). The study context comprised formal conversations in daily work scenarios. Conversations of this type often consist of recurring tasks such as scheduling upcoming meetings. These conversations are concise but require human involvement and a specific degree of cognitive resources (Bastola et al., 2017, p. 2). To validate the hypotheses, an experimental simulation was conducted that aimed to reproduce a daily work-life situation and included a simulated workplace environment. Participants were instructed to complete a complex cognitive task while simultaneously managing meeting schedules in their simulated work calendar and therefore communicating with other artificial employees. Within the chats for scheduling meetings, participants were provided with AI assistance (Bastola et al., 2017,

pp. 3–4). Unlike in the previously outlined studies, participants were not able to oversee and edit generated replies. The design structure of the response assistance followed a dynamic twofold approach. Upon receiving messages, three different response types were generated and shown to the users. These response types were generated by OpenAI's gpt-3.5-turbo model by leveraging the official OpenAI API and including the most recent ten messages of the conversation as context into the prompt. After choosing one response type a second prompt was triggered to generate the actual smart reply given the selected response type. This response message was automatically dispatched to the conversation partner without any opportunity for review or editing (Bastola et al., 2017, p. 4). In contrast to the previously outlined studies and in alignment with the AI-MC classification scheme, this AI-MC approach increases the AI autonomy dimension as users were not able to review messages sent by the system (Hancock et al., 2020, p. 92). Given the variance in the autonomy dimension, it is essential to approach comparisons between the results and findings from different studies with precision. Bastola et al. (2017) first investigated the influence of their AI-based smart reply assistance on overall work performance and mental demand. The quantitative findings suggest that the inclusion of the proposed smart reply assistance in recurring formal conversations enhances overall work performance. In detail, when compared to the control group results, the average number of correct task executions demonstrated an almost significant increase of 18.9%. Productivity, measured in messages per minute, experienced a statistically significant increase of 39% (Bastola et al., 2017, p. 6). Qualitative findings supported these results. Participants agreed that receiving three AI-generated smart replies during chat conversations enhanced productivity in the cognitive task. Despite not directly influencing the task, the autonomously sent replies minimized distraction times, enabling participants returning to the task faster (Bastola et al., 2017, p. 7). Regarding their impact on mental demand, both quantitative and qualitative results indicate that using AI-generated smart replies increased individually perceived task performance, reduced mental demand, and consequently relieved stress (Bastola et al., 2017, pp. 6–7). The statistical analysis of the survey data suggests that the effectiveness of this AI-MC tool in improving task performance influences the willingness to use such a tool in the future (Bastola et al., 2017, p. 9). Despite its beneficial impact on task performance and mental demand, Bastola et al. (2017) also explored drawbacks and proposed future optimizations regarding the design of their AI-MC approach. The primary critical drawback laid in misleading responses of single AI-generated smart reply messages, attributed to a lack of adaption to the context in such instances (Bastola et al., 2017, p. 7). As highlighted in prior studies by Hohenstein and Jung (2018), as well as Mieczkowski et al. (2021), a lack of contextual

relevance was identified as a significant factor for not using Google's smart replies. However, their findings are not directly comparable to the results of the study by Bastola et al. (2017), given the substantial differences in the underlying AI models and, as previously mentioned, the level of autonomy. As direct consequence of the increased AI-autonomy, participants explicitly expressed their desire to have the opportunity to edit the actual reponses sent to their artificial employees (Bastola et al., 2023, p. 7). In contrast to the previously presented AI-MC tool, Fu et al. (2023), much like the study by Inoue et al. (2023), emphasized editable smart replies. They examined the influence of message-level and sentence-level response suggestions on text entry efficiency and perceived usefulness. Expanding upon the outcomes in this context, their objective was to compare the two response assistance options and identify their trade-offs (Fu et al., 2023, p. 1). The contextual focus encompassed political communication calls where purported staffers from legislators were expected to respond to a high amount of citizen inquiries. In the experimental study, the specific task consisted of responding to three email inquiries. Within the task, participants acted as legislator staffers and were provided either with sentence-level, message-level smart replies or no smart replies at all. The smart replies were generated by the GPT-3 model (Fu et al., 2023, pp. 1–3). The quantitative findings revealed that message-level smart replies almost halved the total response time compared to both sentence-level suggestions and the absence of suggestions. Consequently, this indicates that message-level response suggestions generated by the GPT-3 model, enabled faster writing (Fu et al., 2023, p. 5). Fu et al. also examined the editing behavior of participants that used message-level smart replies. Notably, the utilization rate of message-level replies reached 100%. Consistent with results from prior studies, they found that participants mostly edited on top of chosen smart replies. While 25.8% of the selected smart replies were used without change, 73.53% were edited. 1.67% of the chosen replies were entirely deleted, prompting participants to manually rewrite the message. On average, the edited smart replies preserved 75.75% of the tokens initially proposed (Fu et al., 2023, p. 6). This indicates that message-level smart replies generated by the GPT-3 model minimize human-written text as evident in Figure 2: Average percentage of participant-written text per condition (Fu et al., 2023, p. 6).

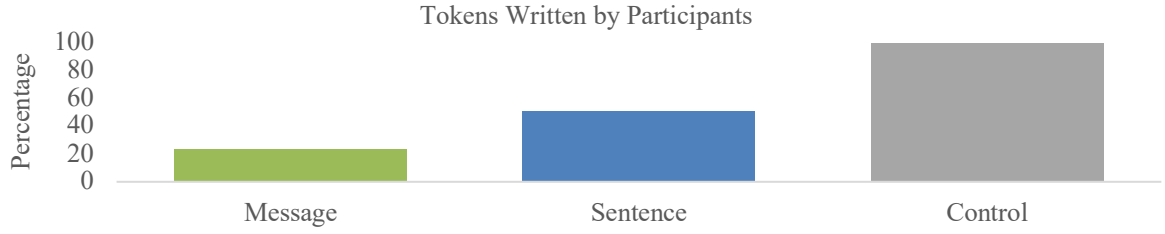

Figure 2: Average percentage of participant-written text per condition (Fu et al., 2023, p. 6)

The significant proportion of edited smart replies, coupled with a 100% utilization rate, underscores the desired ability to edit smart replies, as identified in the previously outlined study by Bastola et al. (2017, p. 7). An analysis of the actual sent responses revealed that those based on message-level suggestions were significantly shorter than human-written responses, averaging only 88 words (Fu et al., 2023, p. 9). This finding contradicts the results of the prior study by Mieczkowski et al. (2021, p. 10), wherein a greater word count was observed following the utilization of smart replies. The study by Fu et al. (2023) also assessed the perceived usefulness of the smart replies by examining perspectives from both writers and readers. Writers mainly perceived message-level smart replies as useful and natural (Fu et al., 2023, p. 6). Their feedback indicates that the message-level drafts were largely perceived as comprehensive response outlines, requiring minimal further adjustments. Participants found them particularly useful in overcoming the initial challenge of ideation when starting to write a full response (Fu et al., 2023, p. 7). From the reader's perspective, message-level smart replies were perceived as most helpful (Fu et al., 2023, p. 9). Concerning the likelihood of future use, participants expressed their willingness to adopt such AI-MC tool in the future. However, they also expressed a discernible hesitancy among participants to entrust its utilization to their legislators. This suggests that in sensitive contexts like political inquiries, the utilization of such a system should be officially disclosed to all communication partners (Fu et al., 2023, p. 8). Fu et al. (2023) also contrasted the beneficial impacts and trade-offs of message-level and sentence-level smart replies. In contrast to message-level smart replies, sentence-level replies did not enable faster response times and were ultimately associated with lower overall satisfaction rates (Fu et al., 2023, pp. 1–2). Fu et al. posited that the reason for the absence of time savings was rooted in the time participants spent deliberating on the candidates for sentence-level suggestions, which offset the time savings from ideation (Fu et al., 2023, p. 5). However, the major trade-off between both assistance options is the level of retained control over the contents of the response message. Message-level smart replies shifted the focus toward editing rather than generating original ideas, as evidenced by the limited contribution of self-written tokens to the message-level suggestions. With sentence-level smart replies, participants acknowledged that they authored the response and were accountable for its contents (Fu et al., 2023, pp. 7–8). In general, the more a smart reply represents a complete response message, the greater the likelihood that users will shift their focus from ideation and drafting to simply editing the given suggestion (Fu et al., 2023, p. 9). As last notably, participants were only provided with one message-level smart reply. Thereupon, some of them expressed their desire to receive more than one option. However, offering multiple response options is closely tied to

increased inefficiency due to the need for decision-making. Such approach may only prove beneficial if the displayed options are complementary (Fu et al., 2023, p. 10).

2.3 Research Hypothesis

This section provided a detailed overview of the state of research in the field of smart replies. Past experimental studies examined their use and identified a range of beneficial impacts. First, with more advanced AI models the overall utilization frequency within the experimental studies increased. While early smart reply systems showed usage rates of only about 6%, smart reply tools based on the GPT model achieved utilization rates of around 32% or even 100%, depending on the study in question (Fu et al., 2023, p. 6; Hohenstein & Jung, 2018, p. 4; Inoue et al., 2023, p. 1975). LSR tools based on the GPT model moreover enabled communication efficiency enhancements as measured in messages per minute or total response times. Following the use of GPT-based LSR, studies for example reported an increase in messages per minute of 18.6% and almost halved total response time (Bastola et al., 2017, p. 6; Fu et al., 2023, p. 5). In this context, the adoption of LSRs based on GPT also decreased response message length compared to the absence of LSRs, emphasizing additional efficiency gains (Fu et al., 2023, p. 9). With regard to individual perception, GPT-based LSRs were mainly perceived as useful and natural across the studies (Fu et al., 2023, p. 6; Mieczkowski et al., 2021, p. 8). However, prior research additionally identified drawbacks and limitations regarding the use of LSRs. First, low utilization rates were primarily grounded in a lack of contextualization within the smart replies (Hohenstein & Jung, 2018, p. 6; Mieczkowski et al., 2021, p. 9). Recent studies integrated context into the prompt by appending the last four to ten messages from the specific chat conversation (Inoue et al., 2023, p. 1975). In such instances, certain LSRs did not align with the ongoing conversation in terms of context (Bastola et al., 2017, p. 7; Inoue et al., 2023, p. 1977). Independent of the model, limited input context is seen as major challenge for the generation of good LSR (Fu et al., 2023, p. 10). Therefore, the primary research gap addressed by this work involves the technical integration of relevant context to generate context aware and helpful LRSs. Additionally, this work will, for the first time, analyze the integration of LSRs within the contextual scope of a job matching environment. Consequently, the research questions (RQ) in this work are derived as follows:

- RQ1: “How can large language models be used to generate reply suggestions within a job matching application using the past chats as context?”
- RQ2: "Does the proposed approach generate well contextualized smart replies?”

- RQ3: “How are the LSRs generated by the proposed approach perceived in a job matching environment?”

Building upon these research questions, the following hypotheses are formulated:

- H1: “It is feasible to design an approach that efficiently incorporates context from past chats to generate LSRs.”
- H2: "The proposed smart reply approach demonstrates proficiency in contextualizing responses for a high conversational relevance."
- H3: “The proposed method to generate LSRs results in high user satisfaction, high perceived naturalness and is likely to be used in the future.”

This work aims to substantiate these hypotheses through a practical research experiment. The subsequent sections will introduce the proposed module for generating contextualized LSRs and then analyze the experiment's results to address the presented research questions.

3 Technical Set-Up of the contextualized LSR module

This section aims to demonstrate how a LLM can be used to generate smart replies within a job matching application while considering past chats. The goals and requirements of the module are outlined, and the applied and implemented LSR model architecture is described, with special emphasis on how the chats are contextualized.

3.1 Goals and Requirements

The main goal of the standalone LSR module is to generate three possible LLM-based smart replies in a chat for the replying user given as context the current and past chats. To make sure that the messages don't sound unnatural, the generated smart replies should adhere to the writer's previous writing style. The smart replies should reply to open points and questions of the chat partner's most recent messages. The smart replies should incorporate already existing knowledge from older chats while not mentioning information and facts which is already mentioned in the current chat. As the generated replies should be further processed and presented so that the user of the model can choose from one of them easily, they should be returned in a common readable format. Furthermore, the replies should be generated in a cost-efficient way. In addition, message-based answer generation should be applied.

3.2 Model Architecture Derivation

As with previous AI-MC research, contextualized LSR are applied, with special emphasis of how the relevant context knowledge is incorporated in an efficient way (Bastola et al., 2017,

p. 4; Inoue et al., 2023, p. 1975) This is especially important, as previous research in AI-MC listed the lack of context knowledge as one of the major challenges of LSR (Bastola et al., 2017, p. 7; Fu et al., 2023, p. 10; Hohenstein & Jung, 2018, p. 6; Mieczkowski et al., 2021, p. 9). The chat data that is incorporated into the LSR can change very rapidly, and many chats might be updated in a short time period. This requires the model to be able to flexibly incorporate dynamic information into the generation of the replies. According to Petroni et al. (2019, p. 2464) language models are able to contain relational knowledge. By fine-tuning the model, additional knowledge can be incorporated. However, a major disadvantage of LLMs is that they are costly to train. Additionally, LLMs only retain training data knowledge, and each subsequent knowledge addition would require retraining the model to integrate it, making it computationally expensive for frequent knowledge updates (Wang et al., 2023, p. 27). Another approach would be to incorporate the knowledge of past chats into the LLM by passing the required context with the smart reply generation prompt, in the way Inoue et al. (2023, p. 1975) and Bastola et al. (2017, pp. 3–4) incorporated the context of the current chat. Instead of only incorporating the current chat context, the past chat context would be passed into the prompt in a similar way. However, prompting has limitations. Firstly, prompting techniques are limited by the number of tokens that can be passed into the LLM. Assuming that the chat example from Appendix 2.3.1 is the currently active chat for which a LSR should be generated. The chat has 281 tokens using the OpenAI Tokenizer (OpenAI, 2024d). Presuming that the prompt takes up 700 tokens, the sum of tokens that we need to be passed to the language model is 981. Given 15 different incorporated chats, the prompt has a size of 5196 tokens, surpassing the token limit of the gpt-3.5-turbo model (OpenAI, 2024b). It is possible to use models with a larger context window and passing in all past context chats, but overall, this leads to higher latency and costs, as OpenAI prices its API based on the tokens passed- and generated by the respective LLM (Liu et al., 2023, p. 10; OpenAI, 2024c). An approach that reduces the prompt size is the Retrieval-Augmented Generation (RAG) model, which provides a framework that combines a text generation component with an information retrieval component that aims to retrieve only the most relevant contextual information for a given input prompt, thus considerably reducing the prompt size (Lewis et al., 2020). As before, together with the prompt, the context information is then passed to a generation component, which generates the desired text while considering the information. In the context of this work's application, there would be no need to pass all the previous chats into the model. Instead, the top-k most relevant previous chat messages from a database are retrieved and then passed into the text generation component together with the prompt. Thus, RAG enables easy access to non-parametric memory, which

can be dynamically updated without fine-tuning. Xu et al. (2023, p. 2) shows that using retriever-augmentation can achieve better performance on long context information than simply passing all the context information into the LLM while being computationally more efficient. Therefore, a RAG-based model is pursued to generate the contextualized LSRs.

3.3 Implemented Model Architecture

Subsequently the RAG information retrieval and generation component are described.

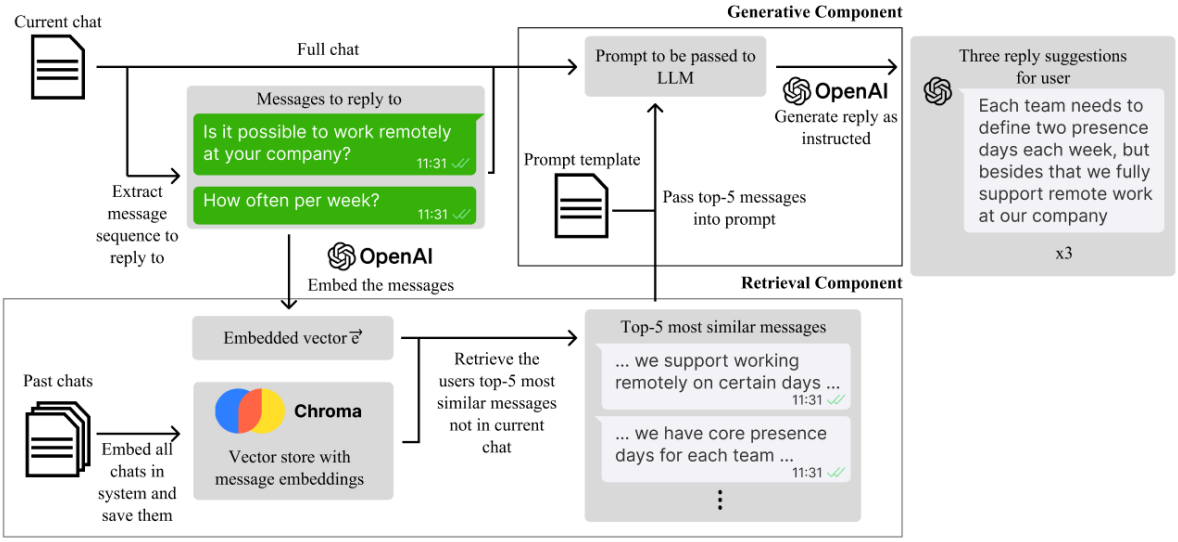

Figure 3: Visualization of the proposed LSR component.

3.3.1 Information Retrieval Component

The context retrieval component is responsible for retrieving the relevant information, which should be passed into the generation prompt for the generation component in the RAG model (See Figure 3, bottom). The term "relevant information" refers to the messages of the replying chat user, which are most similar to the last continuous sequence of messages that the other chat partner sent to the replying user. The goal of the retrieval component is to retrieve the most similar messages from the replying user's old chats. To measure the similarity between the messages to reply to and the old chat messages, embeddings are applied. Embeddings are vectors of decimal numbers which can be dense representations of texts, sentences, and words. Between the two vectors, we compute a similarity metric that reflects how similar the two embedded texts or words are. As recommended by OpenAI we use cosine similarities. The

smaller the similarity metric, the more similar the two embeddings are. The data that should be retrieved can be embedded in different ways (OpenAI, 2024f).

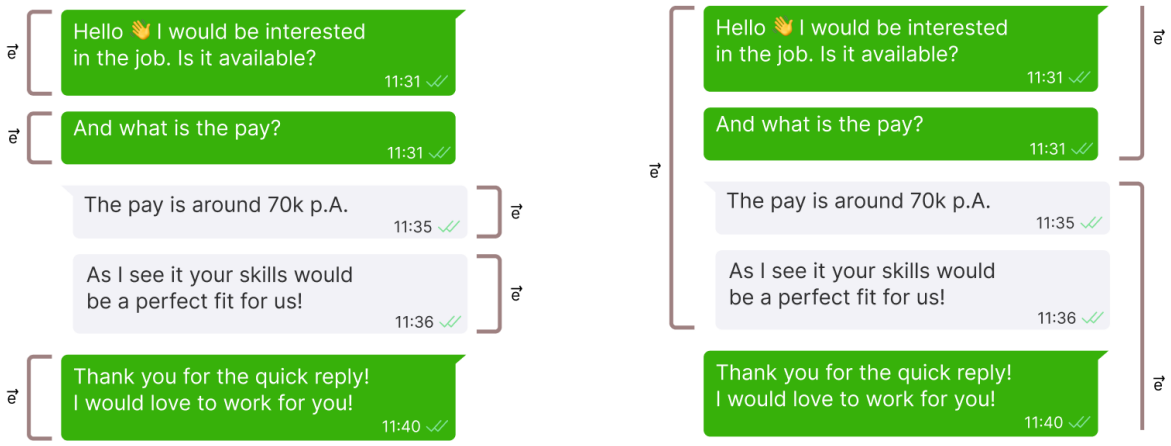

Figure 4: Example for message-based embeddings (left) and message sequence pair embeddings (right) with e denoting each embedding vector.

Two options are considered:

- Each message is embedded on its own, and only the past messages from the user who is currently replying are retrieved. Each message is split up into smaller chunks if the to-be-embedded message is too long (See Figure 4, left illustration).
- Each chat message is embedded as a pair of messages between the chat partners, meaning each contiguous sequence of messages from one chat participant together with each following contiguous sequence of messages from the other chat partner is embedded (See Figure 4, right illustration).

This work adopts the first approach primarily because it ensures that the LSR contains solely the information provided by the responding user. This choice serves to minimize the risk of inadvertently disclosing private information from other users. Additionally, the message sequence pairs would be too long for effective similarity comparisons. During the implementation of the LSR module, it was noticed that the longer an embedded text is, the more likely it is that not only the relevant information is included in one embedding, thus distorting the similarity score. This led to the model not retrieving the relevant information because it would the required embedding had a low similarity score with the message sequence embedding. Smaller chunked message sizes led to better retrieval of similar messages. This issue can't be mitigated for the by chunking the sequence pair embeddings, as the pair-wise information would be lost in subsequent chunks. In conclusion, the data that is queried is

defined as all the replying users' individual chat messages that are not in the current chat. To retrieve the most relevant context, both the to-be-queried data and the search query used to search for the most relevant data need to be embedded. The chat messages that should be replied to are embedded into one single embedding, while the user's chat messages, which are queried for the chat context, are embedded individually. If a chat message is more than 300 characters long it is split into two messages for more fine-grained retrieval of information in long chat messages. The embedded chat messages are then saved in a vector store. A vector store is a database specialized for storing vectors and is optimized for fast retrieval and similarity search (Langchain, 2024). The chat ID and the user who wrote the message are saved in the metadata of each embedded chat. Using OpenAI embeddings, the messages are embedded and stored in the ChromaDB vector store. The retriever component retrieves the top-5 most similar chat messages based on cosine similarities, as recommended by OpenAI (OpenAI, 2024a). This results in a similarity score between zero and one with results closer to one being more similar to each other. Only matches above the threshold of 0.6 are returned, as this shows to work best with the test data. The top-5 messages with the highest similarity score are then passed as context to the generation component.

3.3.2 Generation Component

Due to the time constraints of this work, a pre-trained model is utilized. No LLM will be trained as it requires the needed infrastructure and time to train, which would exceed the time frame given for this work. For the text generation component, OpenAI's GPT LLM is utilized. Namely, the model "gpt-3.5-turbo-0613" is used for testing, and "gpt-4-1106-preview" for the experiment execution. In general, findings from Xu et al. (2023, p. 7) show that it is more effective to choose models with bigger context windows as, as they tend to perform better with RAG with the same passed information. They hypothesize that "... this observation is related to the ‘lost in the middle’ phenomenon ..." (Xu et al., 2023, p. 7), which is reported by Liu et al. (2023, p. 1). Liu et al. (2023, p. 1) show that the performance of a model "... significantly degrades when models must access relevant information in the middle of long contexts ...".Xu et al. (2023, p. 7) provides the example that "... the 4K context LLM tends to ignore the information in the middle of 4K input, while 32K context LLM tend to ignore the information in the middle of 32K input." Thus, "gpt-4-1106-preview" is utilized for the experiments, as the context size receives up to 128.000 tokens (OpenAI, 2024b), promising better incorporation of the passed context information and generation instructions. The model "gpt-3.5-turbo-0613" with a context window of 4096 tokens is used for testing as it is less expensive to use for

frequent testing usage (OpenAI, 2024c). Further considerations include the temperature of the model. The temperature determines how deterministic the LLM chooses the text prediction (Murphy, 2013, p. 103). To ensure that the model acts in a more deterministic way we set the temperature of the model is set to 0.1. With the defined LLM for the experiment, the next step involves instructing it to generate the targeted LSRs. To generate the desired LSRs, prompting techniques are applied to instruct the LLM. To minimize invocations to the LLM, all three messages are generated in one request. Furthermore, aligned with the module requirements, the following instructions and information are included in the prompt in the following order:

- The LLM is instructed with its general task, which it should fulfill, namely, to generate three message-based replies for the user passed into the prompt via the parameter "active_user_id" while considering the current chat history and the provided context.
- A description of the role the replying user takes on is provided in the prompt to better fit the responses into context.
- The LLM is instructed to address all questions and open points of the last message sequence of the chat partner.
- The LLM is instructed to only include information the model knows from the context and the current chat.
- The LLM is instructed that all the replies should convey the same factual content. In early testing it was recognized that the model only incorporates the contextual information in one reply and then makes up the other two generated replies. This instruction enforces, that all replies contain the provided context and chat information.
- The LLM is instructed not to greet the opposing chat partner anymore if the user was already greeted in the chat history, as strangely many generated replies included greetings during testing.
- Under the parameters "previous_chat_context" and "chat_history" the top-5 similar messages which are retrieved from the retrieval component and the full current chat history are passed.
- The message section which should be replied to is passed again, as the model fails to address the latest messages properly without it. This results in better suggestion quality.
- Lastly, the output format instructions are passed below the "format" parameter, which instructs the LLM to return the response in a machine-readable format.

As instructed in the prompt, the retrieved top-5 similar messages, the user id of the replying user, the chat history of the currently active chat and the message sequence to reply to into the generative component are incorporated (See Figure 3, top right). After passing the prompt into

the generative model and retrieving the text completion, the desired smart replies are extracted from the response.

4 Experiment Set-Up

The following section outlines the experiment's setup and provides the rationale behind its design concerning the proposed hypotheses in this study. The experiment is divided into two phases: In the first part, participants are tasked with responding to chat messages within the contextual setting of a job matching application while utilizing the modeled LSR module for assistance. The second part of the experiment requires the participants to fill out a questionnaire to evaluate the given smart replies. Both sections require the participants to input the same unique anonymized name, chosen in advance, to correlate the results of the experimental part and the questionnaire. Participants are selected by recruiting individuals from within the authors' personal network, specifically friends and family members.

4.1 First part: Reply to Chat Messages in a Job Information Setting

In the first part of the experiment, each participant is instructed to successively respond to five messages in a given chat in the role of a given persona. The participant is instructed to use the generated smart replies from the LSR module. Therefore, each participant receives one persona description in advance. In total, there are two personas for job applicants and two for companies. The job seeker personas are designated as "Anna Mueller" and "Emily Taylor," while the company personas go by the names "Greenscape" and "Innovatetech" (see Appendix 9.2). The participants are directed to respond to each chat message to the best of their context knowledge, which is provided in the persona description. The persona descriptions encompass all the necessary details for crafting appropriate responses to the chat messages, and all the knowledge contained in it is represented in past chats the persona had. It is expected that for each experiment chat to be answered, the smart reply module will retrieve the relevant knowledge from the past chats of the persona to generate contextualized smart replies for each experiment chat. To effectively conduct the tests, the LSR module is extended with a user interface, which is presented in Figure 5and a mock chat service, which is responsible for storing and loading the experiment and context chats (See Appendix 2.3). The interface depicts the current chat on top, the three generated LSRs in the middle, and an input field with a "Send" button in which the user can enter the reply to the chat. When a suggestion is clicked on, the text contained is inserted into the input field, overwriting the text in it. The user can then make changes to the text before sending the message. On clicking the "Send" button, the chosen suggestion and the

actual sent response message are saved, and the next chat is automatically loaded. This continues until the participant answers five chats. Based on the selected smart reply and sent text, the BLEU score is calculated to track editing behavior. As each experiment chat requires contextual information from past chats, we can use the BLEU score to deduct how well a message is contextualized. Well-contextualized LSRs should not prompt a perceived need for content editing. Given that the BLEU is 1, no edits were made to the chosen smart reply, and it is assumed that all relevant information from past chats is included, thus implying a high degree of contextualization.

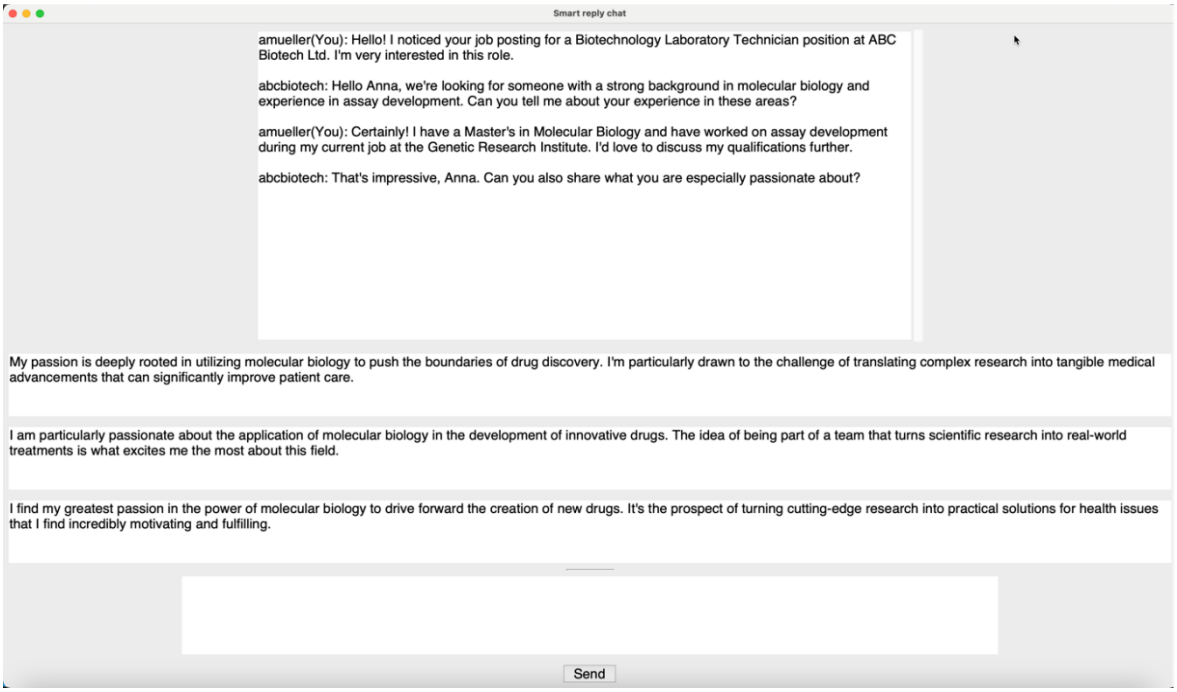

Figure 5: LSR module chat interface excerpt

4.2 Second part: Respond to an Evaluation Questionnaire

The second part of the experiment requires the user to fill out a digital questionnaire on the device the experiment is conducted on. The main questionnaire consists of six closed questions and two open questions. The first step involves participants selecting their age range through a seven-point ordinal scale. This inquiry aims to provide insights into the demographic distribution of participant ages. The remaining closed questions are formulated to align with the goal of verifying hypotheses H2 and H3 of this work. These hypotheses posit that the proposed smart reply approach showcases proficiency in contextualizing responses and leads to high user satisfaction, perceived naturalness, factual correctness, and a strong likelihood of future usage. All of these remaining closed questions are measured by a bipolar 5-point Likert scale. Table 2 lists the remaining closed questions and their reference to each hypothesis.

Hypothesis	Closed questions in the questionnaire
H2	To what extent did the suggestions you selected adequately address the questions and unresolved points raised by your chat partner?
H2	Were the data and information correct and known to you in the generated replies?
H3	How satisfied were you with the proposed text selections?
H3	To what extent, given the previous chat history, did the generated answers sound natural to you?
H3	Considering the context of a job matching application, how likely are you to utilize a smart reply feature?

Table 2: Overview of questionnaire measures mapped to the hypotheses of this work

The findings from the literature analysis indicate that the limited input context is a major challenge regarding the generation of relevant LSRs and therefore lead to irrelevant LSR suggestions in previous experimental studies (Bastola et al., 2017, p. 7; Fu et al., 2023, p. 10; Inoue et al., 2023, p. 1977). Therefore, in addition to the quantitative analysis of the BLEU scores from the first part of the experiment, two additional questions are asked to examine how well the LSR module contextualizes smart replies. Hence, the first two questions shown in Table 2 focus on the perceived contextual coherence and factual correctness of the generated LSRs. The third hypothesis suggests that the proposed LSR model positively impacts human perception of the generated smart replies. Previous study findings already indicate that GPT-3-based LSRs are easy to use, helpful, and useful, leading to high user satisfaction (Fu et al., 2023, p. 7). Consequently, this work directly addresses perceived user satisfaction through the questionnaire. The fourth question aims to confirm that the LLM generates human-like texts, which was already observed in previous studies (Clark et al., 2021, p. 7285; Fu et al., 2023, p. 6; Inoue et al., 2023, p. 1975; Mieczkowski et al., 2021, p. 8). The last question shown in Table 2 seeks to confirm recent findings that people are willing to adopt LSR assistance in the future (Bastola et al., 2017, p. 9; Fu et al., 2023, p. 8). In addition to these five closed questions, two open-ended questions, shown in Table 3, aim to identify areas for future research to enhance the LSR model further.

Category	Open questions in the questionnaire
Strengths	1. What strengths or positive aspects did you notice when using the suggested replies?
Weaknesses	2. In which area do you think the suggested text could be further improved?

Table 3: Summary of the open questionnaire section exploring strengths and areas for improvement.

These questions aim to maintain alignment with previous studies in the field of LSR generation, which have similarly investigated future optimization fields (Bastola et al., 2017, p. 8; Fu et al., 2023, pp. 6–9). All questions, except the question for strengths and weaknesses, are required to be filled in.

4.3 Experiment Process

The experiments are conducted in person on the devices of the authors and are personally accompanied by the experimenters in a distraction-free environment. The experimenters are the authors of this work. The participant is verbally instructed in person by the experimenter and optionally can read over the instructions and experiment details in textualized form. During the whole experiment, the experimenter stays in the room and is available for further questions while not interfering with the participant. After acknowledging the provided information, the experiment starts. The participant is required to read the provided persona description and is then guided to reply to the five provided chats. Consecutively, the participant fills out the questionnaire, and the experiment is concluded. In total, the experiment takes 30 minutes per person and is conducted in English.

4.4 Evaluation Methodology

For the analysis of the closed questions bar- and violin plots are used. For the violin plot, all 5-point Likert-scale answer options are mapped to numerical values ranging from one to five, a higher number meaning higher positive agreement on the posed question. For the bar plots the times an answer option is selected is counted. To evaluate the strengths and weaknesses, a qualitative content analysis approach with inductive categories is applied to find unique statements. In detail, the data is cleaned of incomprehensible answers and grouped by similar content. Answers which contain multiple statements are split up so that each fragment contains one core statement. Finally, categories are defined for each of the groups, with each representing the core statement of the group.

5 Evaluation of the Experiments

This section presents the results of the experiment. The results are subdivided by the hypotheses which are researched in this work and subsequently discussed. In total, the experiment involved 25 participants, with the personas "Greenscape", "Anna Mueller (amueller)", and "Emily Taylor (etaylor)" each comprising six respondents, while "Innovatetech" had seven respondents. All scenarios achieved comparable results with no scenario performing significantly worse or better than others in any researched aspect (See Appendix 2.5). Thus, the focus of the results is on the total aggregated metrics across all scenarios. It needs to be noted that in the qualitative questions querying strengths and weaknesses some participants answered in German. Their answers were translated into English for further analysis.

5.1 Results

5.1.1 Participant Characteristics

The participants of the experiment are predominantly young, with 88% of the participants being younger than 35 years old. The most represented age range is between 18-24, having 16 respondents. Older participants are sparsely represented, with the age range 45-54 having two- and the age range 55-64 having two participants (See Figure 6).

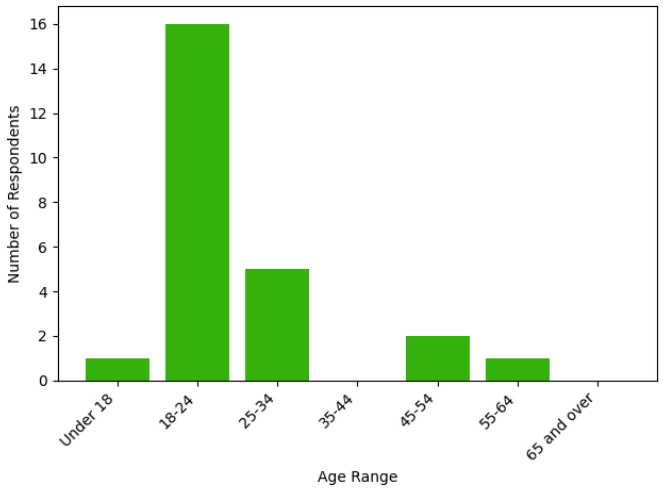

Figure 6: Age distribution among the experiment participants

5.1.2 Contextualization Factors

In this subsection, we present the research characteristics utilized to evaluate the contextualization of the LRS, as defined in sections 4.1 and 4.2. To start, the BLEU score across all answered experiment chats is approximately 0.92, which means that, on average, 92% of the words in the sent responses were generated by the smart reply system. Conversely, approximately 8% of the words in the responses were manually composed by the participants.

Out of 125 messages, 47, or 37.6%, underwent editing, resulting in a BLEU score lower than one. The remaining 62.4% of messages remained unedited. Among the edited messages, the average BLEU score is 0.8, signifying that, on average, 80% of the actual sent and modified response messages were generated by the LSR module introduced in this study.

In regard to factual correctness, 32% of the participants selected that the smart replies are

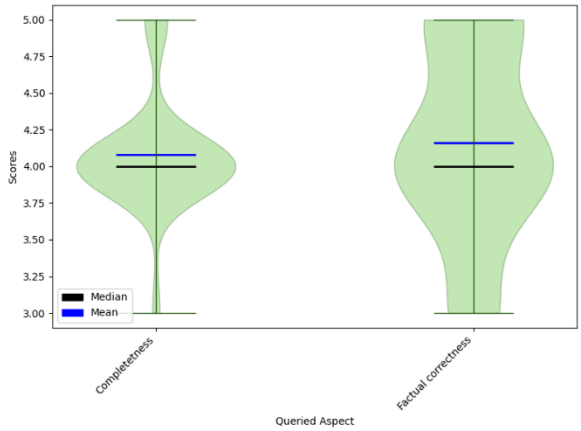

Figure 7: Distribution, mean, and median of numerical mapped answer choices for perceived degree of open point and questions addressed and perceived factual correctness, higher scores indicating agreement

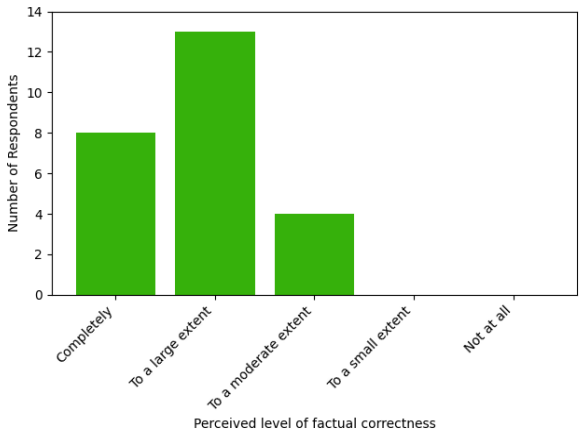

Figure 8: Respondent count per perceived level of factual correctness aggregated per answer choice

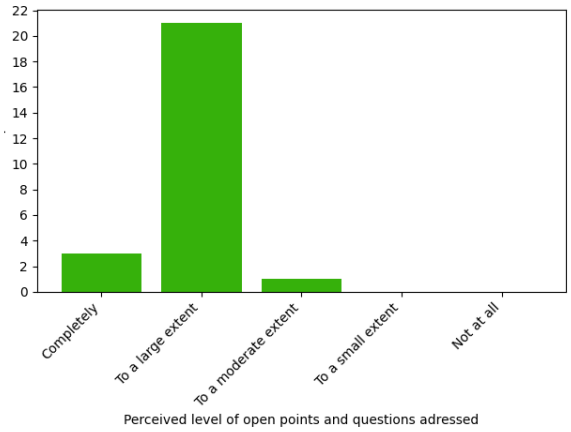

Figure 9: Respondent count per perceived degree of open points and questions addressed aggregated by answer choice

factually "completely" correct and 52% stated that they were correct "to a large extent", while 16% stated that factual correctness is fulfilled "to a moderate extent". No participant reported that factual correctness was only respected "to a small extent" or "not at all" (See Figure 8). In regard to how well the smart replies adequately address open points and questions, 84% of participants responded that they are addressed "to a large extent". 12% stated, that the smart replies "completely" addressed them. As with factual correctness, no participant stated, that the

smart replies were adequately addressing open points and questions only "to a small extent" or "not at all". One participant stated that they are addressed "to a moderate extent" (See Figure 9). Both the perceived factual correctness and the extent to which questions and open points were adequately addressed exhibit means of 4.16 and 4.08, respectively. On average, participants indicated that the LSRs addressed both aspects slightly better than "to a large extent" (See Figure 7). In addition, the qualitative analysis of the experiment revealed that 20% of participants found, that often not all relevant information is be incorporated into the reply. 16% report, that some smart replies are factually incorrect. Furthermore, 16% reported that the smart replies include contextual information and 12% state that the smart replies addressed the questions.

5.1.3 Participant Perception of the Smart Reply Module

Besides researching the capability of the proposed model to generate contextualizes LSR this

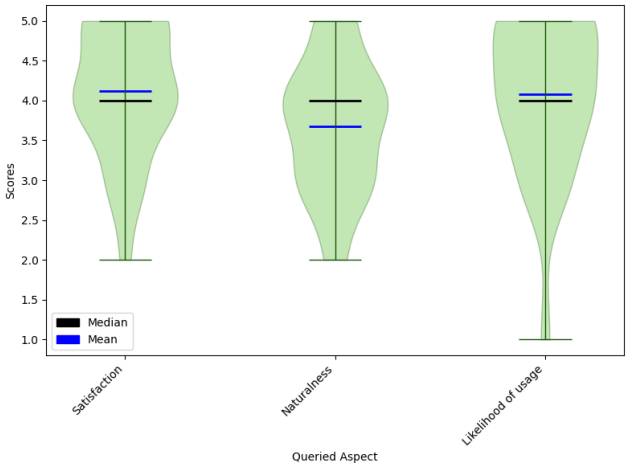

Figure 10: Distribution, mean, and median of numerical mapped answer choices for perceived satisfaction, naturalness of the smart replies and likelihood of usage in the future, higher scores indicating agreement

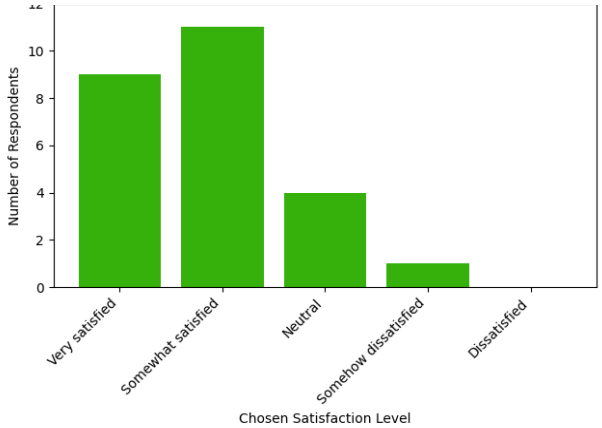

Figure 11: Respondent count per perceived level of satisfaction aggregated by answer choice

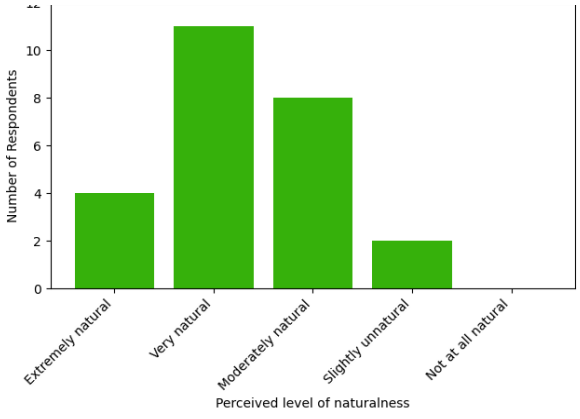

Figure 12: Respondent count per perceived level of naturalness aggregated by answer choice

works aims to investigate the broader perception of the model. In more detail, the overall satisfaction, the naturalness of the smart replies- and the likelihood to use a smart reply tool in the future was investigated. The results show that in each aspect we achieve a high agreement by the participants. In regard to satisfaction, 80% of participants state that they are at least "somewhat satisfied" with the smart replies. In more detail, 44% of participants are "somewhat satisfied" and 36% are "very satisfied" with the suggested replies. One participant reported being "Somehow dissatisfied" by the smart replies while 16% reported to be "neutral" about their satisfaction (See Figure 11). Further, 44% state that the smart replies sounded "very natural". 32% of participants found the smart replies to be "moderately natural". Two respondents found the smart replies to be "slightly unnatural" (See Figure 12). While the mean of the perceived satisfaction is slightly better than "Somewhat satisfied" with a value of 4.12, the mean of the naturalness is slightly below "very natural" with a value of 3.68 (See Figure 10). In terms of likelihood of future usage in a job matching environment, 20% of people are "neutral" about using a smart reply tool, and one participant stated that it is "very unlikely" to

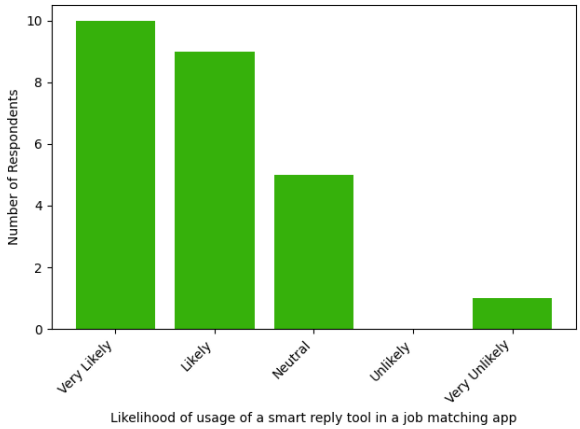

Figure 13: Respondent count aggregated by likelihood of usage

use it. However, 36% of participants would "likely" make use of a smart reply tool in a job matching application and 40% of participants state, that they would "very likely" make use it. In sum, 76% of the participants state that they are at least "likely" to utilize a smart reply system in a job matching environment (See Figure 13).

The qualitative analysis shows shortcomings of the proposed LSR module. 8% of respondents state that the smart replies were not natural enough. 12% state that the replies are too similar and 8% assert that the smart replies contain too much text. The same participant which stated that he very unlikely will utilize a smart reply module stated that "[...] a trained human will be just as fast in writing the text themself instead of reading/adapting the generated one". However, 16% of participants stated that the smart replies sounded natural and 8% claimed that

the smart replies sounded professional. One participant proclaims that the replies "[...] are very easy to read and use." and one states that he is satisfied by the smart replies. Another participant asserts, that the smart replies helped to "start somewhere". Overall, 12% of participants state that the smart replies helped them save time.

5.2 Discussion

Aligned with hypothesis H2 of this study, a key interest was to evaluate whether the proposed LSR module exhibits a high proficiency in contextualizing LSR suggestions. The findings of the experiment reveal that 62.4% of the sent replies were not edited before being sent. Given that the content of each experiment chat was existing within past chats, it is indicated that these 62.4% of the LSR suggestions encompassed a highly appropriate level of relevant contextual information derived from the personas' past interactions. Directly compared to the past study conducted by Fu et al. (2023, p. 6), who observed that 25.8% of the suggestions were used without editing, this implies a higher and better degree of contextualization. his conclusion gains additional support when comparing the average BLEU score observed in this study with the most comparable studies by Fu et al. (2023, p. 6) and Inoue et al. (2023, p. 1977). They reported BLEU scores of around 0.75 and 0.62 respectively. With 92%, the average BLEU score in this study, calculated over all replies, is 22.67% and 48.39% higher, respectively. However, even if restricting the analysis towards all used but edited LSR suggestions, with on average 80% of the words remaining unchanged between the LSR suggestions and the actual sent response message, it is indicated that that even if changes needed to be made, they were relatively small. As outlined in Chapters 4.1 and 4.2 of this study, for a comprehensive verification of hypothesis H2, alongside the examination of BLEU scores, the perceived factual correctness, and the effectiveness in addressing questions and open points by the LSRs were also assessed. As evident in Figure 7, both measures were rated high, meaning that the majority of the provided LSR suggestions contained relevant contextual information to formulate an appropriate response. Hence, the results of this study confirm that the proposed LSR module demonstrates proficiency in contextualizing LSR generations based on past chats and therefore confirm the second hypothesis of this work. The second key outcome of this study refers to user perception of the LSR suggestions in order to verify hypothesis H3 of this work. Throughout the experiment, participants reported high overall user satisfaction and high naturalness regarding the LSR selections. Moreover, they mainly agreed that they are likely to use a contextualized LSR tool in a job matching setting in the future. Hence, as expected all three measures underscored results from previous studies (Bastola et al., 2017, p. 9; Clark et al., 2021,

p. 7285; Fu et al., 2023, 6; Inoue et al., 2023, p. 1975; Mieczkowski et al., 2021, p. 8). Therefore, hypothesis H3 is considered confirmed.

6 Implications, Limitations and Future Research

6.1 Implications

The successful integration of contextual knowledge from past interactions positions the developed model as a promising tool for optimizing communication in job matching applications. While this study may not have employed a large, representative sample size, it gives an outline how well contextualized LSRs can be implemented and evaluated in the job matching domain. User satisfaction and likelihood of future utilization of the model underscore its potential for future adoption in the industry. In practice the proposed LSR module could improve communication between job seekers and employers, providing quick and relevant responses that enhance the overall usability of the platform. Further, as implied by three participants in our study, the module can lead to time and resource efficiency for both job seekers and recruiters. Quick and accurate responses facilitate a more efficient communication process, reducing the time spent on repetitive tasks. For companies, it could speed up the recruitment- and application process, and job seekers and talents might find their desired jobs faster than in the past. In addition, the usage of a LSR module may attract more people to the job matching platform, as users are likely to make use of such a tool. Further, the study contributes to the research landscape by addressing gaps in contextualization within AI-generated smart replies, providing valuable insights for future developments in the field.

6.2 Limitations

The experiment and the questionnaire were conducted in English and mainly questioned non-native English speakers. Furthermore, the sample size of 25 participants is small and mainly consisted of friends and relatives of the authors. Thus, the experiment data was not sampled directly from the target group of users of job matching applications. making the experiment not representative. Furthermore, findings from the quantitative and qualitative questionnaire show, that the contextualization doesn't always work well. Participants reported that the smart replies did not always include all information and were not always factually incorrect. This was reported by respectively 20% and 16% of participants. It can be assumed that in such cases, the LSR module struggles to retrieve the relevant information and, in some cases, only returns a fraction- or even entirely wrong information. Another possible reason is that the underlying LLM hallucinates false facts. In addition, while the naturalness of the smart replies is perceived

to be high, it is the lowest rated aspect of the researched LSR module. We hypothesize this can be accredited to the fact that the smart replies are generated based on contextual knowledge from artificially created context- and experiment chats. The model is instructed to adhere to the user's writing style when generating the smart replies, thus incorporating the writing style of the artificial chats into the smart replies, which do not necessarily reflect real conversations in a job matching scenario. In turn, this can make the smart replies be perceived as less natural. In addition, the conditions of the experiment might not reflect the same conditions as a real-life chat application in regard to scale. The experiment- and past chats are relatively small and few in number. In real-life applications, users might have many more and more length chats, which would need to be considered, which could lead to different results.

6.3 Future research

The development of the LSR module showed several opportunities for future research. Firstly, the module's temporal sense can be researched. For example, given a chat message which was written a year ago which states "I would be free to start working next month.". If the user would now be asked when he could start working, the technical module could deem the stated message as similar and include it in the context. The temporal information that the user can start next month is not updated based on the date it was written, which could lead to the generation of factually wrong smart replies. Related is the research on the depreciation of old information in the vector store. While it is possible to filter the context chat messages on retrieval with the date the chat messages were written on, there can be more nuanced ways to determine when old data should be deleted. For example, old vector store entries could be deprecated when the user did not use the application in a long time, indicating that their background information might have changed. To add to that, a mechanism to filter factually identical messages can be introduced, as ensure diversity in the retrieved top-k messages. Furthermore, the context integration of the job seeker- and company profile or the companies' job posting can be tackled to generate better contextualized smart replies. Language inclusivity is another area for exploration, as expanding the module's support beyond English would make the LSR module more accessible. In future works the possibilities of integrating different media types such as pictures, videos and voice messages can be studied. Assessing scalability with an increased number of chats and users and more lengthy conversations is important for understanding the real-world applicability of the module. Additionally, fine-tuning the model's parameters to achieve better optimized and more contextually aware results is a key focus for future research. Furthermore, with the regard to the experiment, a follow-up study can be conducted which samples the participants from a

predefined target group while also including native English speakers to better evaluate smart reply quality whilst ensuring representability. Additionally, further research can be performed on how the contextualized LSR performs on other act types as reported by Inoue et al. (2023, p. 1977).

7 Conclusion

This seminar thesis aimed to develop and evaluate a model for generating contextualized smart replies in job matching applications by incorporating knowledge from past chats. The following key findings have been shown: Firstly, a model for efficiently generating contextualized smart replies in a job matching environment was outlined and developed (See RQ1). Secondly, it is shown that the outlined model is able to successfully integrate contextual knowledge from past chats into the smart replies (See RQ2). Lastly, it is demonstrated that the users of the model are satisfied by the generated smart replies, perceive them as natural and would likely use such a tool in the context of a job matching application (See RQ3). The last two key findings have been derived by conducting experiments in which participants received a defined persona for which they needed to reply to five chats. The participants were instructed to utilize the smart reply model and, if needed, to edit the smart reply before submitting the reply. All chats needed contextual information from past chats to be answered, which was stated in the persona the participants received. The experiment application tracked how much the participants altered the chosen smart reply. After concluding the experiment, the participants evaluated the proposed model. While the proposed model is already able to successfully generate contextualized smart replies on the small experiment chat dataset, it is recommended that before applying it in a real-world scenario further research is conducted on how the model performs and scales in real-life scenarios.

1 References

Arnold, K. C., Chauncey, K., & Gajos, K. Z. (2020). Predictive Text Encourages Predictable Writing. In F. Paternò, N. Oliver, C. Conati, L. D. Spano, & N. Tintarev (Eds.), *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 128–138). ACM. https://doi.org/10.1145/3377325.3377523

Bastola, A., Wang, H., Hembree, J., Yadav, P., McNeese, N., & Razi, A. (2017). LLM-based Smart Reply (LSR): Enhancing Collaborative Performance with ChatGPT-mediated Smart Reply System. In J. Tenenberg, J. Sheard, D. Chinn, & L. Malmi (Eds.), *Proceedings of the 2017 ACM Conference on International Computing Education Research* (pp. 1–11). ACM. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Chairs), *Proceedings of the 34th International Conference on Neural Information Processing Systems.* Symposium conducted at the meeting of Curran Associate Inc.

Buschek, D., Zürn, M., & Eiband, M. (2021). The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In Y. Kitamura, A. Quigley, K. Isbister, & T. Igarashi (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). ACM. https://doi.org/10.1145/3411764.3445372

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings ofthe 59th Annual Meeting ofthe Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 7282–7296). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2107.00061

Fu, L., Newman, B., Jakesch, M., & Kreps, S. (2023). Comparing Sentence-Level Suggestions to Message-Level Suggestions in AI-Mediated Communication. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on*

Human Factors in Computing Systems (pp. 1–13). ACM. https://doi.org/10.1145/3544548.3581351

Goldenthal, E., Park, J., Liu, S. X., Mieczkowski, H., & Hancock, J. T. (2021). Not All AI are Equal: Exploring the Accessibility of AI-Mediated Communication Technology. *Computers in Human Behavior,*(125), 1–9. https://doi.org/10.1016/j.chb.2021.106975

Google (Ed.). (2023, November 14). *Google Allo has signed off.* https://allo.google.com/

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication,*(25), 89–100. https://doi.org/10.1093/jcmc/zmz022

Hohenstein, J., & Jung, M. (2018). AI-Supported Messaging. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Eds.), *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). ACM. https://doi.org/10.1145/3170427.3188487.

Inoue, M. M., Tomoyuki Shibata, & Ryoichi. (2023). The Effect of Response Suggestion on Dialogue Flow: Analysis Based on Dialogue Act and Initiative. In M. Goldwater, F. Anggoro, B. Hayes, & D. Ong (Eds.), *Proceedings ofthe 45th Annual Conference ofthe Cognitive Science* (45th ed.).

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)* (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300469

Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., Corrado, G., Lukacs, L., Ganea, M., Young, P., & Ramavajjala, V. (2016). Smart Reply: Automated Response Suggestion for Email. In B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1–10). ACM. https://doi.org/10.1145/2939672.2939801

Koubaa, A., Boulila, W., Ghouti, L., Alzahem, A., & Latif, S. (2023). Exploring ChatGPT Capabilities and Limitations: A Survey. *IEEE Access*, *11*, 118698–118721. https://doi.org/10.1109/ACCESS.2023.3326474

Langchain. (2024, January 18). *Vector stores*. https://python.langchain.com/docs/modules/data_connection/vectorstores/

leap in time. (2023, August 7). *Lithire*. https://www.leap-in-time.com/services/lithire/?lang=de

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, & H.-T. Lin (Chairs), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*. http://arxiv.org/pdf/2005.11401.pdf

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023, July 6). *Lost in the Middle: How Language Models Use Long Contexts*. http://arxiv.org/pdf/2307.03172.pdf

McKinsey & Company. (2023, August 25). *What's the future of generative AI? An early view in 15 charts*. https://www.mckinsey.com/featured-insights/mckinsey-explainers/whats-the-future-of-generative-ai-an-early-view-in-15-charts#/

Mieczkowski, H., Hancock, J. T., Naaman, M., Jung, M., & Hohenstein, J. (2021). AI-Mediated Communication. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–14. https://doi.org/10.1145/3449091

Murphy, K. P. (2013). *Machine learning: A probabilistic perspective* (4. print. (fixed many typos)). *Adaptive computation and machine learning series*. MIT Press.

OpenAI. (2024a, January 10). *OpenAI Platform Embeddings - Limitations & risks*. https://platform.openai.com/docs/guides/embeddings/limitations-risks

OpenAI. (2024b, January 10). *OpenAI Platform Models*. https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

OpenAI. (2024c, January 10). *OpenAI Platform Pricing*. https://openai.com/pricing#language-models

OpenAI. (2024d, January 10). *OpenAI Platform Tokenizer*. https://platform.openai.com/tokenizer

OpenAI. (2024e, January 17). *OpenAI Platform*. https://platform.openai.com/docs/models/gpt-3-5

OpenAI. (2024f, January 18). *OpenAI Platform - Embeddings*. https://platform.openai.com/docs/guides/embeddings

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Chairs),

Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. https://aclanthology.org/p02-1040.pdf

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language Models as Knowledge Bases? In S. Padó & R. Huang (Chairs), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. https://aclanthology.org/D19-1250/

PWC. (2017). *Sizing the prize: PwC's Global Artificial Intelligence Study: Exploiting the AI Revolution*. https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf

Reuters Media. (2023, February 2). *ChatGPT sets record for fastest-growing user base - analyst note*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

SACKS, H., SCHEGLOFF, E. A., & JEFFERSON, G. (1974). A Simplest Systematics for the Organization of Turn Taking for Conversation**This chapter is a variant version of "A Simplest Systematics for the Organization of Turn-Taking for Conversation," which was printed in Language, 50, 4 (1974), pp. 696–735. An earlier version of this paper was presented at the conference on "Sociology of Language and Theory of Speech Acts," held at the Centre for Interdisciplinary Research of the University of Bielefeld, Germany. We thank Dr. Anita Pomerantz and Mr. Richard Faumann for pointing out to us a number of errors in the text. *Studies in the Organization of Conversational Interaction*, 696–735. https://doi.org/10.1016/B978-0-12-623550-0.50008-2

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, *67*, 1–24. https://doi.org/10.1016/j.csl.2020.101151

Wang, S., Zhu, Y., Liu, H., Zheng, Z., Chen, C., & Li, J. (2023, October 24). *Knowledge Editing for Large Language Models: A Survey*. http://arxiv.org/pdf/2310.16218.pdf

Xu, P., Ping, W., Wu, X., McAfee, L., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoeybi, M., & Catanzaro, B. (2023). *Retrieval meets Long Context Large Language Models*. http://arxiv.org/pdf/2310.03025.pdf

2 Appendix

2.1 Tables

Dimension	Definition	Examples
magnitude	The extent of the changes that AI enacts on messages	Correcting spelling errors vs. generating entirely newmessages
media type	The media in which AI operates (e.g., text, audio, video)	Suggesting text replies vs. Modifying one's appearance in video
optimization goal	The goal for which AI is optimizing the messages	To appear attractive, trustworthy, humorous, dominant, etc.
autonomy	The degree to which AI can operate on messages without the sender's supervision	Sender chooses between AI suggested messages vs. AI engages in conversation with minimal input from the sender
role orientation	Therolethatthe AI isoperating on behalfof (e.g., sender vs. receiver)	Sender: offering messages to enhance reply efficiency vs. Receiver: assessing whether sender is potentially lying

Table 4: Overview of the AI-MC tool classification dimensions

2.2 Persona Summaries

2.2.1 Summary of the Persona Description for Anna Müller (amueller)

"Anna Müller, a Molecular Biologist at the Genetic Research Institute, is a leader in assay development. With a Master's in Molecular Biology, she successfully led a team in creating a groundbreaking drug discovery assay ahead of schedule. Her proficiency spans PCR, DNA sequencing, and collaborative project coordination. Anna's strengths lie not only in technical expertise but also in her commitment to teamwork, open communication, and proactive learning. Currently Anna is searching for a new job as she is not satisfied with her current salary. Ready for a new role from February 2024, Anna brings a dynamic blend of skills and passion to contribute to innovative projects in molecular biology. She has a salary expectation of 70.000 to 80.000 dollars."

2.2.2 Summary of the Persona Description for Emily Taylor (etaylor)

“Emily Taylor, thriving at Digital Dynamics Inc., is a strategic sales expert. She achieved a 20% revenue boost through targeted market strategies, excelling in high-value account management. Emma's efficiency initiatives, like implementing a new CRM system, led to a substantial 20% increase in client satisfaction scores and business growth. Known for proactive problem-solving, transparent communication, and alignment with company objectives, Emma's adaptability and continuous learning make her a valuable asset in the dynamic intersection of sales and technology. Currently looking for a new job to face new challenges in her professional life.”

2.2.3 Summary of the Persona Description for GreenScape (GreenScape)

“GreenScape is a dedicated urban planning and architecture studio, which is committed to environmental consciousness and sustainability. Prioritizing collaboration and innovative projects, exemplified by initiatives like Omega GreenConnect, the studio fosters a culture of environmental innovation and close collaboration with local communities, focusing on modern and eco-friendly planning practices. GreenScape offers diverse career paths in urban planning and architecture, including comprehensive employee benefits and mentorship programs for individual growth. With upcoming projects in the revitalization of urban spaces, the studio positions itself at the forefront of sustainable practices, environmental awareness, and impactful innovation. GreenScape is currently seeking professionals for key urban planning roles such as Urban Design Strategist, Environmental Impact Analyst, and Sustainable Development Planner.”

2.2.4 Summary of the Persona Description for InnovateTech (InnovateTech):

“InnovateTech is a dynamic consulting firm which helps its customers in adapting the newest innovations in their business. While the company has broad capabilities in the area of innovation management, its main focus is on helping customers adapt AI and Blockchain technologies. The company is dedicated to fostering a thriving remote work culture, emphasizing a healthy work-life balance. Leveraging various collaboration tools and virtual meetings, they prioritize open communication and values such as proactive communication, adaptability, and accountability. The company places a strong emphasis on team dynamics, organizing virtual team-building activities to create a cohesive remote team. Professional growth is a cornerstone, with cross-functional projects, mentorship programs, and tailored learning plans. The company offers various career growth paths and ensures equal

representation of men and women in the board area. With a focus on AI and blockchain, InnovateTech is at the forefront of technology, driving transformation in their customers busines operations.”

2.3 Chat Mock Service

Before starting the development of the module, a mock service that retrieves chats and users who are active in these chats is established. The data structure of the mock service consists of two objects: The chat and the user profile. The chat has a unique ID and stores the participating users and chat messages. Each chat message contains a time stamp, text content, and the sender of the message. For each user, the chats in which they are active, what role they have, and a unique identifier is saved. Around this data structure, we implement a wrapper class that can get the chats and users by their ID.

2.3.1 Mocked Chat


```
{
 "chatId": 1,
 "participatingUsers": ["amueller", "univrecruiter"],
 "messages": [
  {
   "senderUserId": "amueller",
   "content": "Hello, I came across the Computer Science Laboratory Technician position at XYZ University. I'm reaching out to express my interest in this role.",
   "timestamp": "2023-01-15T14:30:00"
  },
  {
   "senderUserId": "univrecruiter",
   "content": "Hello Anna, we're looking for someone with a strong background in computer science and experience in software development. Can you provide more details about your experience in these areas?",
   "timestamp": "2023-01-15T14:35:00"
  },
  {
```

"senderUserId": "amueller",
"content": "Certainly! I hold a Master's in Computer Science and have hands-on experience in software development during my current position at the University's Computer Science Lab. I would be happy to discuss my qualifications further.",
"timestamp": "2023-01-15T14:40:00"
},
{
"senderUserId": "univrecruiter",
"content": "That's acceptable, Anna. Additionally, could you shed light on any challenges you've faced in computer science and software development, and how you've overcome them?",
"timestamp": "2023-01-15T14:45:00"
}
]
}

2.3.2 Mocked profile

{
"id": "etaylor",
"role": "job_seeker",
"activeChats": ["9", "10", "11", "12", "13", "31", "32", "33", "34", "35"]
}

2.4 Prompts

2.4.1 Company prompt

Please generate three possible chat replies for user {active_user_id}, taking into account the chat so far and especially focusing on incorporating the knowledge from the context.

User {active_user_id} is a representative of a company and is seeking people to hire for a job. The context includes past interactions and discussions that occurred between the company {active_user_id} and potential job seekers. The replies should address all questions and open points of the latest message from the job seeker.

The replies should only include relevant information known through the context messages from previous chats and is needed to answer the text. Avoid making up new information and don't deviate from the user's writing style. Also, please ensure that all replies convey the same information but are rephrased differently. Don't greet the user which you are replying to again if you already greeted him in the chat history.

Context:

{previous_chat_context}

Chat history:

{chat_history}

Latest message from user {chat_partner_user_id} to reply to:

{message_to_reply_to}

{format}

2.4.2 Job seeker prompt

Please generate three possible chat replies for user {active_user_id}, taking into account the chat so far and especially focusing on incorporating the knowledge from the context.

User {active_user_id} is a job seeker and is writing with a representative from a company which offers a job he is interested in. The context includes past interactions and discussions that occurred between the job seeker {active_user_id} and companies he is interested in. The replies should address all questions and open points of the latest message from the company representative.

The replies should only include relevant information known through the context messages from previous chats and is needed to answer the text. Avoid making up new information and don't deviate from the user's writing style. Also, please ensure that all replies convey the same

information but are rephrased differently. Don't greet the user which you are replying to again if you already greeted him in the chat history.

Context:

{previous_chat_context}

Chat history:

{chat_history}

Latest message from user {chat_partner_user_id} to reply to:

{message_to_reply_to}

{format}

2.5 Additional Visualizations and Statistics by Scenario

Scenario	Average percentage of words in the replies that come from the human participant
Amueller	11.19 +- 11.32
Etaylor	4.6 +- 3.6
Greenscape	4.56 +- 7.72
Innovatetech	7.87 +- 7.08
Total	**7.53 +- 7.08**

Table 5: Average percentage of words in the replies that come from the human participant per scenario

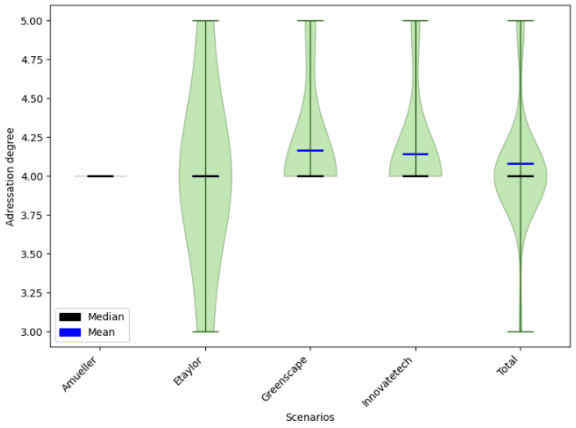

Figure 14: Distribution of perception on how well questions and open points are addressed by the smart replies by scenario

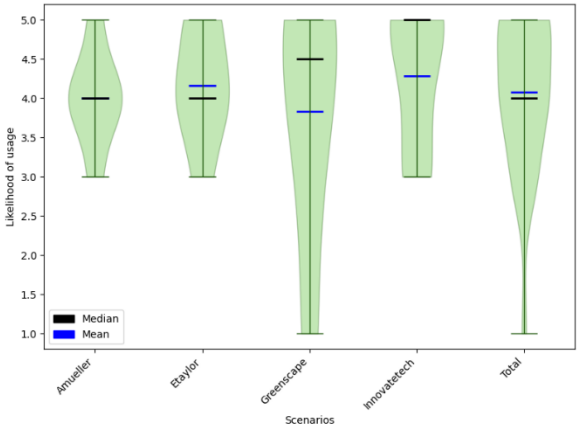

Figure 15: Distribution of likelihood of usage of a smart reply system in a job matching application by scenario

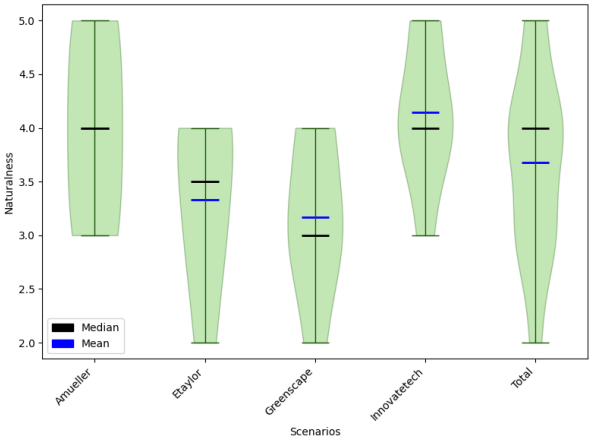

Figure 16: Distribution of perceived naturalness of the smart replies by scenario

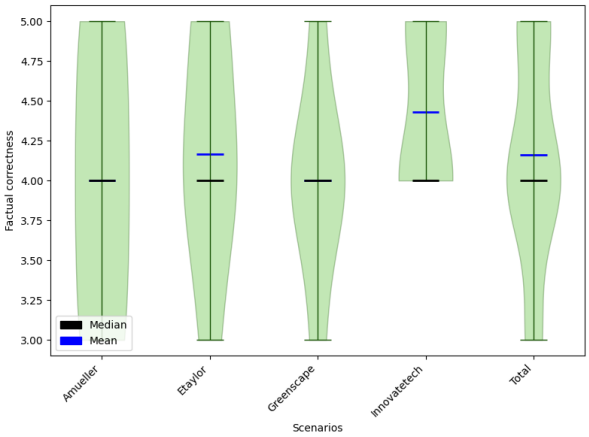

Figure 17: Distribution of perceived factual correctness by scenario

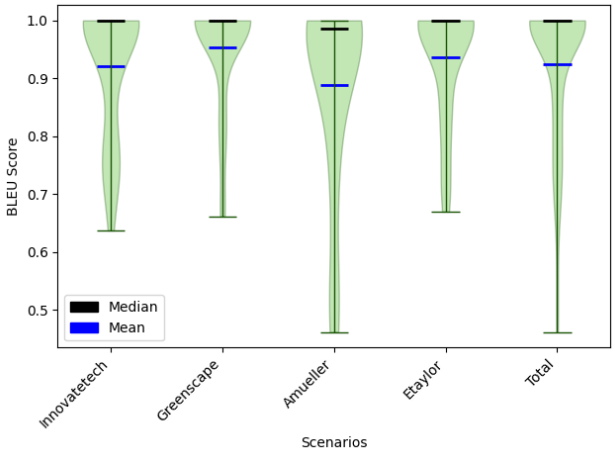

Figure 18: Distribution of BLEU scores by scenario

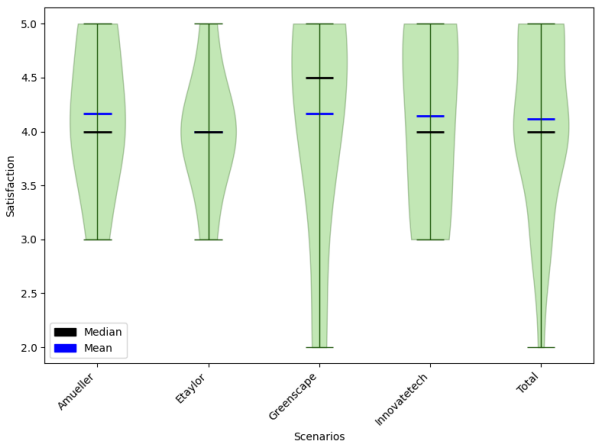

Figure 19: Distribution satisfaction with the LSR tool per scenario