Examining Lexical Alignment in Human-Agent Conversations with GPT-3.5 and GPT-4 Models

BOXUAN WANG, University of Twente, The Netherlands

This study employs a quantitative approach to investigate lexical alignment in human-agent interactions involving GPT-3.5 and GPT-4 language models. The research examines alignment performances across different conversational contexts and compares the performance of the two models. The findings highlight the significant improvements in GPT-4's ability to foster lexical alignment, and the influence of conversation topics on alignment patterns. By providing insights into these aspects, this research aims to contribute to the development of more engaging and effective conversational agents.

Additional Keywords: lexical alignment, human-agent interaction, GPT-3.5, GPT-4

1 INTRODUCTION

The growing significance of conversational agents in daily life, supported by their ongoing evolution, has led to a wide range of applications, including but not limited to agents functioning as healthcare helpers, customer service assistants, learning guides, and emotional companions [3,9,11,13]. Due to this increasing involvement of agents in human life, there is an increasing need to not only advance AI technologies but also gain a profound understanding of the factors contributing to successful human-computer interactions.

Linguistic alignment, a notion first systematically accounted for by Pickering and Garrod [14], refers to the process where two speakers in a conversation adjust to each other's linguistic behaviors to be more aligned in the representations of what is being communicated. This phenomenon, according to Pickering and Garrod [14,15], can be activated on multiple levels, including phonological, lexical, syntactic, and semantic alignment. Among these, lexical alignment, pertaining to the adoption of the same lexical items [14], has piqued broad research interest. Studies have demonstrated that lexical alignment can result in heightened engagement and rapport, as well as successful accomplishment of tasks among human-human interlocutors [2,16].

In the scope of human-agent interaction, investigating lexical alignment is essential in understanding its influence on both human users and conversational agents. Insights can be derived regarding how agents can better adapt to users linguistically and provide more engaging and efficient conversations, resulting in more satisfied interactions for the users [7,17]. Furthermore, examining the metrics and patterns of alignment enables the identification and assessment of the areas where agents excel or fall short in aligning with users, which in turn can aid the design, development, and optimization of agents.

TScIT 39, July 7, 2023, Enschede, The Netherlands

Recent months have witnessed a breakthrough in highly sophisticated large language models, among which GPT models, developed by OpenAI, have demonstrated state-of-the-art competencies in natural language generation and comprehension. Specifically, GPT-3.5 and GPT-4, as the latest versions of the GPT model, have showcased significant technological progress in terms of handling complex language tasks and engaging with users. The current study aims to examine lexical alignment in human-agent conversations by concentrating on GPT-3.5 and GPT-4 models. By delving into the lexical alignment patterns and comparing their performances, the study seeks to contribute to the understanding of the implications of advancements of large language models on human-agent communication and provide information on the development of more engaging and effective conversational agents.

2 PROBLEM STATEMENT

Despite existing research on lexical alignment in the realm of human-agent interactions, many of these studies have employed meticulously programmed rule-based agents [10,17] or Wizard-of-Oz systems [1,5,7,8], controlling various degrees of lexical alignment as independent variables to investigate user satisfaction or task completion performance. While these studies offer valuable insights, they do not directly examine the extent to which state-of-the-art large language models, such as GPT-3.5 and GPT-4, align lexically with human users. As these advanced models become increasingly accessible, understanding their lexical alignment with human users is crucial for developing more engaging and effective conversational agents. Furthermore, comparing the alignment performances of GPT-3.5 and GPT-4 may reveal the impact of advancements in large language models on human-agent interactions.

However, the exploration of lexical alignment involving these models remains limited. This study aims to address this gap by analyzing lexical alignments in human-agent conversations, specifically with GPT-3.5 and GPT-4 models. The first research question guiding this study is:

RQ1: How does lexical alignment in conversations with human participants differ between GPT-3.5 and GPT-4?

In addition to the comparison between different versions of the GPT model, the influence of conversation topics on lexical alignment is also of interest. Different conversation topics may elicit different patterns of language use and alignment. Therefore, the second research question is:

RQ2: How does lexical alignment in conversations between GPT models and human participants differ between task-oriented and non-task-oriented topics?

3 RELATED WORK

Lexical alignment, as proposed by Pickering and Garrod [14,15], is one aspect of the broader concept of alignment in conversation, which can occur at various levels, including phonological, lexical, syntactic, and semantic. In this framework, lexical alignment specifically refers to the phenomenon where interlocutors start to use the same words or phrases during a conversation. This alignment is achieved by a "priming mechanism", which refers to the activation of a particular linguistic representation, such as a word or a phrase, making it more likely that the representation will be reused. Another notion is "routinization", a form of priming where interlocutors develop and rely on shared routines, which are mutually agreed upon ways of expression in a certain conversation that drastically reduce the cognitive effort of language production and comprehension. Pickering and Garrod [14,15]'s model offers valuable insights into the mechanism of efficient conversation and laid the groundwork for later theoretical and empirical studies of lexical alignment in human-human and human-agent conversations.

Among the empirical studies in human-agent conversations, Koulouri et al. [7] investigated lexical alignment in a Wizard-of-Oz human-agent conversation involving a visual task, focusing on its occurrence and reciprocity, revealing the role of lexical alignment in stabilizing the vocabulary employed. Importantly, they found that lower alignment correlated with less successful interactions. Similarly, Spillner and Wenig [17] investigated linguistic alignment in an information retrieval task with a carefully crafted chatbot that can adjust levels of both lexical and syntactic alignment. The results revealed that employing lexical and syntactic alignment can reduce user workload and increase user engagement.

Duplessis et al. [1] proposed a framework to quantify lexical alignment and self-repetition behaviours based on a sequential pattern mining approach, and further conducted a comparative study of human-human and human-agent lexical alignment based on corpora of task-oriented conversations. Using the framework, they discovered that human-human conversation showcased more flexibility in alignment, and both parties' behaviours in human-human conversations are more homogenous than those in human-agent conversations.

Although there is a lack of research on alignment directly involving GPT-3.5 and GPT-4 models as they are only recently made available, studies [6,12] have highlighted the similarities and differences between the two models in terms of architecture, training process, and performance. Koubaa [6] noted that GPT-4 retained the same transformer-based architecture of GPT-3.5 but with a significant expansion in model size and the incorporation of a rule-based reward model to fine-tune its performance. OpenAI [12] reported that GPT-4 outperformed its predecessors in various benchmark tests, including language tests designed for humans, and demonstrated considerable enhancement in its ability to follow user intent.

Given their exposure to human-generated text during training, it is reasonable to expect that both GPT-3.5 and GPT-4 might have been implicitly trained to align with the lexical choices of their conversation partners to some extent to facilitate communication. Furthermore, the substantial improvements in GPT-4's architecture, training process, and performance suggest that it might exhibit different patterns of lexical alignment compared to GPT-3.5.

4 METHODOLOGY

4.1 Research design

The research adopted a 2x2 factorial experimental design using two variables: the GPT model version (GPT-3.5 or GPT-4.0) and the conversation topic (task-oriented topic or non-task-oriented topic). The topics were predetermined: a collaborative storytelling task (task-oriented) and a casual conversation about hobbies (non-task-oriented). The collaborative storytelling task involves participants and the GPT models creating a story together by taking turns contributing sentences, while also allowing for conversations between participants and the models to discuss and shape the story they are creating. The casual conversation about hobbies, on the other hand, is less structured and more open-ended, allowing participants and the GPT models to freely discuss their interests and experiences related to various hobbies.

The rationale behind choosing these predetermined topics was twofold. Firstly, it was intended to account for the diverse applications of GPT models. Unlike previous studies with rule-based or Wizard-of-Oz agents that are typically designed for a particular experiment, GPT models are versatile and can be used in various contexts. By choosing different topics, the study sought to encompass a broader range of use cases of GPT models, thereby providing a more comprehensive understanding of their alignment performance across different contexts. Secondly, previous research has suggested that the alignment patterns may differ between task-oriented and non-task-oriented conversations, with greater divergence expected in casual conversations [4]. Examining this difference is an important aspect of understanding lexical alignment in human-agent conversations [1].

4.2 Data collection

Conversational data for the experiment were collected from 20 participants proficient in English, who were recruited through personal contacts. Each participant engaged in a conversation with both GPT-3.5 and GPT-4 models. To ensure variety in their interactions, each participant was assigned one of the two predetermined topics for their conversation with GPT-3.5 and the other topic for their conversation with GPT-4. This process led to a balanced assignment of topics to GPT models, resulting in an equal number of conversations per topic per model. The order of interactions was randomized to control for any potential order effects.

Each conversation lasted for 15 turns, which was determined based on the results of pilot studies. This number of turns allowed participants to maintain active engagement without fatigue and produced sufficient conversational data for the analysis of lexical alignment. Prompts were designed to initiate the conversations and were provided to the participants at the beginning of each conversation. The duration for each participant to complete both conversations was around 25-40 minutes. All conversations took place through the online interface provided by OpenAI, the developer of GPT models, and were automatically captured by the interface in the form of transcripts.

Participants provided informed consent for the use of their conversational data in the study and were explicitly instructed not to disclose personal identifiable information during the conversations. They were also informed that OpenAI would also have access to the conversational data.

Table 1. Summary of metrics and corresponding descriptions

Metric	Description
EV (Expression Variety)	Proportion of unique shared expressions relative to all tokens in a conversation
ER (Expression Repetition)	Proportion of repetitions of shared expressions relative to all tokens in a conversation
VO (Vocabulary Overlap)	Proportion of overlapping tokens relative to all tokens in a conversation
ENTR (Entropy)	Shannon entropy of the length of shared expressions
L (Average Length)	Average length of the shared expressions
LMAX (Maximum Length)	Maximum length of the shared expressions
IE_S (Initiated Expressions)	Proportion of shared expressions initiated by the speaker relative to all shared expressions
ER_S (Expression Repetition)	Proportion of repetitions of shared expressions relative to all tokens by the speaker
$Tokens_S$	Proportion of tokens produced by the speaker relative to all tokens in a conversation
VO_S (Vocabulary Overlap)	Proportion of overlapping tokens relative to all tokens produced by the speaker
SEV_S, SER_S, $SENTR_S$, SL_S, $SLMAX_S$	Speaker-specific version of corresponding metrics that focus on self-repetition lexicon

4.3 Data preprocessing

Data preprocessing involved several steps in preparing the collected conversational data for the analysis stage. The first step was the correction of typos in participants' inputs. This decision was justified since the intended words could be reliably inferred based on orthographic similarities, the context of the typos, and the responses from the GPT models, which could correctly detect and interpret these mistakes. The motivation behind this step was to represent the intentions of the participants more accurately in the transcripts, allowing for a more precise assessment of lexical alignment.

Then, the data was tokenized and normalized using the NLP library spaCy, with customization in the handling of contractions, capitalization, and punctuation. Afterwards, the tokenized and normalized data were converted into a tab-separated values (.tsv) file, making the data compatible with the alignment analysis tool used in the next stage of the study.

4.4 Data analysis

Analysis of the conversational data in this study is guided by the framework established by Duplessis et al. [1]. This framework offers a comprehensive and structured approach for quantifying lexical alignment, proposing a range of metrics that quantify both speaker-independent and speaker-dependent aspects of the conversational data. To implement this framework, a tool provided by Duplessis et al. was used. This tool takes a corpus of tab-separated values (.tsv) files as input and generates results for the various alignment metrics. Table 1 provides a summary of these metrics.

The speaker-independent metrics (EV, ER, VO, ENTR, L, LMAX) assess the overall conversation, focusing on the shared lexicon between speakers. The shared lexicon refers to the set of shared expressions in a conversation. A shared expression refers to a string of tokens that occurs in utterances made by both speakers, and at least once in a "free form", which means that the expression is not syntactically dependent on another segment of the utterance such as being part of a larger expression. This requirement ensures that the shared expressions are distinct and meaningful units that independently contribute to the conversation at least once.

The speaker-independent metrics can be divided into two groups based on what they measure: the usage of shared expressions (EV, ER, VO) and the characteristics of the shared expressions themselves (ENTR, L, LMAX). In the first group, EV measures the number of unique shared expressions normalized by the total number of tokens, capturing the variety of alignment process. ER quantifies the frequency of repetitions of shared expressions, reflecting the strength of repetition of lexical alignment. VO, a broader and more inclusive metric, calculates the proportion of overlapping tokens relative to all tokens produced in the conversation, without distinguishing between unique or repeated expressions or considering the "free form" requirement. In the second group, ENTR measures the complexity of shared expressions in terms of their lengths using Shannon entropy, a concept from information theory that quantifies the unpredictability or randomness of information. L and LMAX measure the average and maximum length of shared expressions respectively.

The speaker-dependent metrics, on the other hand, are divided into two groups: those that examine each speaker's interaction with the shared lexicon and their overall contribution to the conversation (IE_S, ER_S, $Tokens_S$, VO_S), and those that focus on the speaker's self-repetition behaviors (SEV_S, SER_S, $SENTR_S$, SL_S, $SLMAX_S$). Among these, ER_S and VO_S are speaker-specific versions of ER and VO respectively. The self-repetition behaviors are analyzed by looking at the speaker's self-expression lexicon, which refers to the set of expressions that a speaker uses more than once, thereby reflecting the repetition of their own lexical choices.

5 RESULTS

This section presents the findings of the study, addressing the research questions outlined earlier. The analysis focuses on the lexical alignment in conversations between human participants and two versions of the GPT model (GPT-3.5 and GPT-4), and how it varies between the topics of collaborative storytelling and casual conversation about hobbies. The findings are presented in two parts: first is the discussion of the role of the GPT model version, followed by an examination of the role of the topic.

Table 2. Average values and standard deviations of descriptive statistics for each sub-corpus

	Tokens	Shared Lexicon Size
GPT-3.5 Story	1839.9±453.9	170.8±41.8
GPT-3.5 Casual	1376.8±188.4	133.4±30.1
GPT-4 Story	1575.8±242.7	184.4±28.5
GPT-4 Casual	1175.6±222.9	144.5±33.3

The data for this study consists of 40 conversations, each containing 30 utterances (15 utterances by each speaker). These conversations are divided into four sub-corpora of 10 conversations based on the 2x2 factorial design, with each

Table 3. Average values and standard deviations of speaker-independent metrics

	GPT-3.5			GPT-4		
	Storytelling	Casual	Combined	Storytelling	Casual	Combined
EV	.093±.011	.096±.011	.095±.011	.117±.005	.123±.015	.120±.011
ER	.430±.032	.395±.038	.412±.039	.459±.021	.422±.034	.441±.033
VO	.203±.029	.244±.038	.223±.039	.247±.031	.298±.037	.272±.042
ENTR	1.171±.114	.869±.133	1.020±.196	1.127±.164	.887±.146	1.007±.195
L	1.396±.070	1.252±.062	1.324±.098	1.377±.096	1.258±.059	1.318±.099
LMAX	5.200±1.549	6.100±2.025	5.650±1.814	6.600±5.168	5.500±1.354	6.050±3.720

Table 4. Contrastive comparisons of GPT-3.5 and GPT-4 based on speaker-independent metrics

GPT-3.5 vs GPT-4	**EV**	**ER**	**VO**	**ENTR**	**L**	**LMAX**
Storytelling	<***	=	<**	=	=	=
Casual	<**	=	<*	=	=	=
Combined	<***	<*	<**	=	=	=

Statistical difference was determined using Wilcoxon rank-sum tests. Asterisks denote different levels of p-values (*p<0.05, **p<0.01, ***p<0.001); "=" signifies that the observed difference is not statistically significant (p≥0.05).

sub-corpus representing a specific condition (GPT-3.5 or GPT-4, storytelling or casual conversation). Table 2 presents descriptive statistics of each sub-corpus, including the number of tokens and the size of the shared lexicon.

As can be seen, the number of tokens and the size of the shared lexicon vary across the different conditions. Notably, the GPT-3.5 model tends to produce a higher number of tokens than the GPT-4 model, and the storytelling topic generally results in a higher number of tokens than the casual conversation topic. In terms of the size of the shared lexicon, the GPT-4 model and the storytelling topic both tend to have a larger shared lexicon size than their counterparts.

5.1 Role of GPT model version

The aim of this section is to identify significant differences in lexical alignment behavior that can be attributed to the version of the GPT model used in the conversation, so the focus is on the combined dataset that includes both the storytelling and casual conversation topics for each model. For completeness, results from the individual topics are also presented.

5.1.1 Speaker-independent metrics. Speaker-independent metrics provide an overview of the lexical alignment in the conversations as a whole. While these metrics do not differentiate between the contributions of the human participant and the GPT model, the assumption is that both the GPT models (GPT-3.5 and GPT-4) and their human conversation partners are relatively consistent in their behavior across different conversations. Therefore, any significant differences observed in these metrics between the GPT-3.5 and GPT-4 conversations can be attributed to the differences in the models' behavior.

Table 3 presents the average values and standard deviations of the speaker-independent metrics for both GPT-3.5 and GPT-4. The contrastive comparisons of these metrics between GPT-3.5 and GPT-4 are further detailed in Table 4. As shown in the tables, there are significant differences between the two versions of the GPT model; specifically, GPT-4 exhibits a higher EV, ER, and VO compared to GPT-3.5. On the other hand, there are no statistically significant differences in ENTR, L, and LMAX.

EV, which measures the size of the shared lexicon normalized by the length of the conversation, is higher in conversations with GPT-4 compared to conversations with GPT-3.5 (see Figure 1). This suggests that unique shared expressions are established more frequently in conversations with GPT-4. This could imply that GPT-4 is more effective at using a diverse vocabulary that aligns with the human participant's language, so that a higher percentage of its vocabulary is reciprocated by the participant. Alternatively, it could suggest that GPT-4 is better at adapting its language to the conversation, leading to a higher percentage of the participant's lexical choices being reciprocated by GPT-4. Both scenarios contribute to a larger shared lexicon.

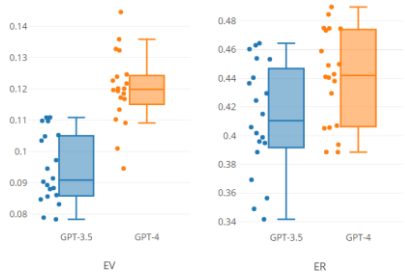

Fig. 1. Comparison of EV and ER between GPT-3.5 and GPT-4

ER, which measures the percentage of tokens speakers use as repetitions of shared expressions, is higher in conversations with GPT-4 (see Figure 1). This suggests that shared expressions are repeated more frequently within a conversation with GPT-4. This could indicate that GPT-4 is more adept at fostering such repetition, either by reusing the same expressions itself to maintain coherence or reinforce certain points, or by influencing the human participant to increase their repetition of certain expressions. The difference in ER between GPT-3.5 and GPT-4 is statistically significant only when considering the combined data of both topics, but not in each individual sub-corpus. This could suggest that the sample size for each sub-corpus is too small to detect the difference, but when the data from both topics are aggregated, the difference becomes more apparent.

VO, which measures the proportion of overlapping tokens out of all tokens, provides a more general sense of lexical

Table 5. Average values and standard deviations of speaker-dependent metrics

	GPT-3.5			GPT-4		
	Storytelling	Casual	Combined	Storytelling	Casual	Combined
IE$_S$	.498±.073	.429±.051	.463±.071	.491±.070	.504±.054	.497±.061
ER$_S$	.412±.047	.401±.053	.406±.049	.451±.053	.404±.058	.427±.059
Tokens$_S$	.656±.098	.719±.055	.687±.083	.499±.042	.512±.092	.505±.070
VO$_S$	.264±.059	.296±.048	.280±.055	.382±.054	.447±.073	.415±.071
SEV$_S$	.159±.009	.181±.009	.170±.014	.155±.013	.172±.015	.163±.016
SER$_S$	.677±.036	.711±.025	.694±.035	.632±.028	.613±.048	.622±.039
SENTR$_S$	1.432±.168	1.372±.125	1.402±.147	1.124±.203	1.032±.124	1.078±.170
SL$_S$	1.568±.111	1.511±.077	1.539±.098	1.377±.104	1.324±.056	1.350±.086
SLMAX$_S$	6.200±1.932	5.400±.699	5.800±1.473	3.900±.876	4.800±1.619	4.350±1.348

Table 6. Contrastive comparisons of GPT-3.5 and GPT-4 based on speaker-dependent metrics

GPT-3.5 vs GPT-4	**IE$_S$**	**ER$_S$**	**Tokens$_S$**	**VO$_S$**	**SEV$_S$**	**SER$_S$**	**SENTR$_S$**	**SL$_S$**	**SLMAX$_S$**
Storytelling	=	=	>***	<***	=	>*	>**	>**	>**
Casual	<***	=	>***	<***	=	>***	>***	>***	>*
Combined	=	=	>***	<***	=	>***	>***	>***	>***

Statistical difference was determined using Wilcoxon rank-sum tests. Asterisks denote different levels of p-values (*p<0.05, **p<0.01, ***p<0.001); "=" signifies that the observed difference is not statistically significant (p≥0.05).

Table 7. (A)symmetry between GPT-3.5 and human speakers based on speaker-dependent metrics

GPT-3.5 vs Human	**IE$_S$**	**ER$_S$**	**Tokens$_S$**	**VO$_S$**	**SEV$_S$**	**SER$_S$**	**SENTR$_S$**	**SL$_S$**	**SLMAX$_S$**
Storytelling	=	<*	>***	<***	=	=	>*	>*	=
Casual	<***	=	>***	<***	>**	>***	>***	>***	>**
Combined	<**	=	>***	<***	=	>***	>***	>***	>**

Statistical difference was determined using Wilcoxon rank-sum tests. Asterisks denote different levels of p-values (*p<0.05, **p<0.01, ***p<0.001); "=" signifies that the observed difference is not statistically significant (p≥0.05).

Table 8. (A)symmetry between GPT-4 and human speakers based on speaker-dependent metrics

GPT-4 vs Human	**IE$_S$**	**ER$_S$**	**Tokens$_S$**	**VO$_S$**	**SEV$_S$**	**SER$_S$**	**SENTR$_S$**	**SL$_S$**	**SLMAX$_S$**
Storytelling	=	=	=	=	=	=	=	=	<**
Casual	=	=	=	=	=	=	=	=	=
Combined	=	=	=	=	=	=	=	=	=

Statistical difference was determined using Wilcoxon rank-sum tests. Asterisks denote different levels of p-values (*p<0.05, **p<0.01, ***p<0.001); "=" signifies that the observed difference is not statistically significant (p≥0.05).

alignment, without distinguishing between unique or repeated expressions. A higher VO in conversations with GPT-4 suggests that a larger proportion of the tokens used in the conversation are overlapping tokens. This could indicate that GPT-4 is more adept at fostering alignment with the human participant, leading to a larger overlap in the words and phrases used by both parties.

EV, ER and VO each offer a unique perspective on the usage of aligned expressions. They are not always in sync; for instance, a conversation could have a wide variety of unique shared expressions (high EV) but these expressions might not be repeated often (low ER), or vice versa. The fact that all three metrics are significantly higher in conversations with GPT-4 suggests a more robust engagement with the shared lexicon by one or both speakers compared to conversations with GPT-3.5. In contrast, the lack of statistically significant differences in ENTR, L, and LMAX between GPT-3.5 and GPT-4 suggests that the shared expressions in conversations with the two models are similar in terms of their complexity, average length, and maximum length.

These findings suggest that the improvements in GPT-4 are primarily related to the level of engagement with the shared lexicon, rather than to the complexity or length of the shared expressions used. In other words, while the shared lexicon in conversations with GPT-4 tends to be more often established and more frequently repeated, the complexity and length of these shared expressions themselves do not significantly differ from those in conversations with GPT-3.5.

5.1.2 Speaker-dependent metrics. The speaker-dependent metrics provide a more detailed look at the lexical alignment behavior of the GPT models by differentiating between the behaviors of the human participant and the GPT model. These metrics can be divided into two groups: those that measure the shared lexicon and overall contribution to the conversation (IE$_S$, ER$_S$, Tokens$_S$, and VO$_S$), and those that measure the self-repetition lexicon (SEV$_S$, SER$_S$, SENTR$_S$, SL$_S$, SLMAX$_S$). Table 5 presents the average values and standard deviations of these metrics.

The metrics related to shared lexicon provide insights into how the GPT models align their language with the human participant; the self-repetition lexicon metrics shed light on the models' behavior regarding repeating their own expressions, which can be instrumental in understanding their

lexical alignment behavior with human participants. Each group of metrics is examined in two ways. First, the metrics are compared between GPT-3.5 and GPT-4 to identify any significant differences in the models' behavior, as shown in Table 6. However, to fully understand these differences, it's important to consider them in the context of human behavior. Therefore, the metrics are also compared between the GPT models and their respective human conversation partners to further investigate any (a)symmetry in their lexical alignment behaviors, as shown in Table 7 and Table 8. For example, if a model exhibits a behavior that is more similar to humans, it could be interpreted as an improvement in that aspect of lexical alignment.

In the group of shared lexicon related metrics, the first metric to consider is IE_S, which measures the ratio of shared expressions initiated by a speaker. When examining the IE_S between GPT-3.5 and GPT-4, there is no significant difference (see Figure 2). When comparing the GPT models with human participants, it is found that GPT-3.5 initiates fewer shared expressions than humans. In contrast, GPT-4 demonstrates a level of symmetry with human participants, initiating shared expressions as frequently as they do (see Figure 2). This observation suggests that GPT-4 has improved its ability to contribute to the shared lexicon in a conversation by initiating expressions that the human participant is likely to pick up and reuse. It's also worth noting that the difference in IE_S between the models is less pronounced than the difference between the models and human participants, which could be due to the conversation topic, the size of the corpus, or the statistical nature of these comparisons.

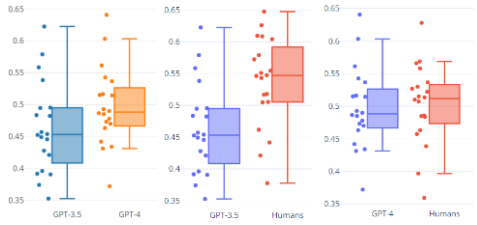

Fig. 2. Comparison of IE_S between GPT-3.5 and GPT-4, and each model vs their respective human partners

In the analysis of ER_S, which is the speaker-dependent version of ER that measures the percentage of a speaker's tokens as repetitions of shared expressions, no significant difference is found between GPT-3.5 and GPT-4. This suggests that both models exhibit a similar frequency of using shared expressions in their own utterances. When comparing the GPT models with their human conversation partners, symmetry is observed in both cases, indicating that the models and their human partners use shared expressions in their utterances at similar rates. It is worth noting that GPT-3.5 already demonstrated a level of ER_S comparable to human partners, which may account for the lack of significant improvement observed in GPT-4 in this regard. Interestingly, a discrepancy arises when comparing these results with the speaker-independent metric ER, as examined in Section 5.1.1. It was found that conversations involving GPT-4 exhibit a higher frequency of shared expression repetition than those with GPT-3.5. This discrepancy between ER and ER_S could suggest that the difference in ER may be influenced more by GPT-4's ability to enhance the human participant's use of shared expressions, either by generating expressions that are subsequently picked up and reused by humans, or by validating and encouraging the human participant's use of certain expressions, leading to their increased repetition. The former interpretation aligns with the results shown by the IE_S metric.

For $Tokens_S$, which measures the proportion of tokens produced by a speaker in the conversation, GPT-3.5 produces more tokens overall than GPT-4 (see Figure 3). This is further supported by the results that GPT-3.5 exhibits asymmetry with human partners, producing more tokens than them, while GPT-4 shows symmetry, producing a similar number of tokens as human partners. For VO_S, which measures the proportion of a speaker's tokens that overlap with human partners', GPT-4 has a higher VO_S than GPT-3.5 (see Figure 3). Moreover, GPT-4 shows symmetry with human partners in terms of VO_S, while GPT-3.5 exhibits a smaller VO_S than human partners. When considering these two metrics together, a noteworthy pattern emerges. Although GPT-4 contributes a smaller proportion of the total tokens in a conversation compared to GPT-3.5, a larger proportion of its tokens overlap with human partners. This suggests that GPT-4 may be more selective or efficient in its use of language, indicating a higher level of lexical alignment with the human participant.

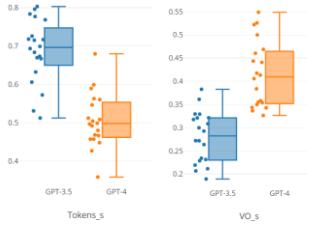

Fig. 3. Comparison of $Tokens_S$ and VO_S between GPT-3.5 and GPT-4

Fig. 4. Comparisons of SEV_S, SER_S, $SENTR_S$, SL_S, and $SLMAX_S$ between GPT-3.5 and human partners

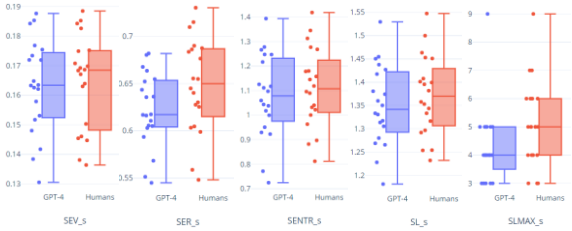

Fig. 5. Comparisons of SEV_S, SER_S, $SENTR_S$, SL_S, and $SLMAX_S$ between GPT-4 and human partners

In the group of self-repetition lexicon related metrics, GPT-3.5 and GPT-4 show no significant difference in SEV_S, which measures the variety of self-repeated expressions. This

Table 9. Average values and standard deviations of speaker-independent metrics

	Storytelling			Casual Conversation		
	GPT-3.5	GPT-4	Combined	GPT-3.5	GPT-4	Combined
EV	.093±.011	.117±.005	.105±.015	.096±.011	.123±.015	.109±.019
ER	.430±.032	.459±.021	.444±.030	.395±.038	.422±.034	.409±.038
VO	.203±.029	.247±.031	.225±.037	.244±.038	.298±.037	.271±.045
ENTR	1.171±.114	1.127±.164	1.149±.139	.869±.133	.887±.146	0.878±.136
L	1.396±.070	1.377±.096	1.387±.082	1.252±.062	1.258±.059	1.255±.059
LMAX	5.200±1.549	6.600±5.168	5.900±3.782	6.100±2.025	5.500±1.354	5.800±1.704

Table 10. Contrastive comparisons of storytelling topic and casual conversation topic based on speaker-independent metrics

Story vs Casual	EV	ER	VO	ENTR	L	LMAX
GPT-3.5	=	=	<**	>***	>***	=
GPT-4	=	>*	<*	>**	>**	=
Combined	=	>**	<**	>***	>***	=

Statistical difference was determined using Wilcoxon rank-sum tests. Asterisks denote different levels of p-values (*p<0.05, **p<0.01, ***p<0.001); "=" signifies that the observed difference is not statistically significant (p≥0.05).

suggests that both models exhibit a similar level of diversity in their self-repetitions. However, for the other metrics - SER_S, $SENTR_S$, SL_S, and $SLMAX_S$, which measure the frequency, entropy of length, average length, and maximum length of self-repeated expressions respectively - GPT-3.5 scores higher than GPT-4. This indicates that GPT-3.5 tends to repeat its own expressions more frequently, with a higher level of complexity and length.

Similarly, when comparing the GPT models with their human conversation partners, GPT-3.5 exhibits asymmetry, scoring higher than humans in SER_S, $SENTR_S$, SL_S, and $SLMAX_S$ (see Figure 4). This suggests that GPT-3.5's self-repetition behavior is more pronounced than that of humans in terms of frequency, complexity, and length. On the other hand, GPT-4 shows symmetry with human partners in these metrics (see Figure 5), indicating that its self-repetition behavior is more aligned with that of humans.

Overall, the speaker-dependent metrics reveal a consistent pattern of symmetry between GPT-4 and humans across all measures, including metrics related to both shared lexicon and self-repetition lexicon. This symmetry, which is not observed with GPT-3.5, suggests that GPT-4 has adopted a more human-like conversational strategy. This strategy include improvements such as initiating expressions that are likely to be picked up and reused by human partners, producing fewer tokens but with a larger proportion that overlap with human partners, and in its self-repetition behavior. While not all of these strategies strictly represent GPT-4's lexical alignment behavior in the traditional sense, they do facilitate a shared lexicon and foster a conversational environment that encourages more lexical alignment from the human partners, contributing to a more natural and engaging interaction.

5.2 Role of conversation topic

While the previous section focused on differences attributable to the version of the GPT model, the aim of this section is to identify significant differences in lexical alignment behavior that can be attributed to the conversation topic. The focus is on the combined dataset that includes both GPT-3.5 and GPT-4 for each conversation topic. For completeness, results from the individual models are also presented. The analysis is limited to speaker-independent measures, as the conversations for each topic include both GPT models as speakers, making it less suitable to use speaker-dependent metrics, which are designed to measure the behavior of a single speaker. This approach allows us to isolate and assess the impact of conversation topics on lexical alignment, highlighting their role independent of speaker-specific behaviors.

Table 9 shows the average values and standard deviations of the speaker-independent metrics for both the storytelling and casual conversation topics. Table 10 further details the comparisons of these metrics between the two topics. As shown in the tables, there are significant differences between the two conversation topics. Specifically, the storytelling topic exhibits a higher ER, ENTR, and L compared to the casual conversation topic; on the other hand, the casual conversation topic has a higher VO.

The ER metric, which measures the proportion of total tokens that are repetitions of shared expressions, is higher for the storytelling topic than for the casual conversation topic (see Figure 6). This suggests that in storytelling, the conversation tends to feature more frequent reuse of shared expressions. This is probably due to the narrative nature of storytelling, where certain characters, events, or themes are repeatedly mentioned to maintain coherence and continuity in the story. On the other hand, the VO metric, which measures the overlap in vocabulary between the two speakers, is higher for the casual conversation topic (see Figure 6). This could be because casual conversations often involve common topics and everyday language, leading to a higher degree of vocabulary overlap.

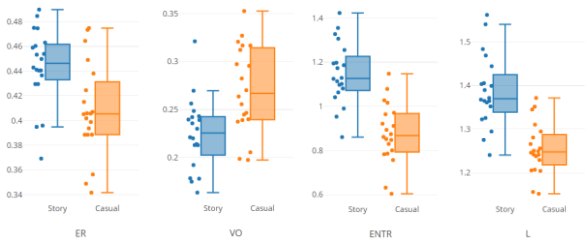

Fig. 6. Comparison of ER, VO, ENTR and L between the topic of collaborative storytelling and the topic of casual conversation

As for the ENTR and L metrics, which measure the complexity and average length of shared expressions respectively, both are higher for the storytelling topic (see Figure 6). This suggests that the shared expressions used in storytelling are both longer and more complex than those used in casual conversations. This could be due to the descriptive and detailed nature of storytelling, which often requires the use of longer and more complex shared expressions to convey the story effectively and maintain continuity.

The analysis of the role of the conversation topic reveals that the nature of the conversation can indeed influence lexical alignment patterns. Specifically, the storytelling topic exhibits higher levels of ER, ENTR and L, showing greater lexical alignment. However, VO is higher in casual conversations, indicating a larger overlap in the vocabulary used by both parties. These findings underscore the importance of considering the conversation topic when analyzing and interpreting lexical alignment patterns in conversations with AI agents.

6 DISCUSSION

6.1 Major findings

The results highlight the significant role of the GPT model version in shaping the pattern of lexical alignment in conversations. Notably, GPT-4 demonstrated a better ability to foster lexical alignment compared to GPT-3.5, as evidenced by its superior performance in several aspects, such as the variety of expressions introduced into the shared lexicon (EV) and the efficient use of language to align with human partners (VO, Tokens$_S$, VO$_S$). Specifically, GPT-4 exhibits a stronger ability to initiate expressions that are subsequently picked up (IE$_S$) and reused (ER, ER$_S$) by human partners, indicating a more proactive role of GPT-4 in facilitating lexical alignment by making it easier for humans to align with the model. However, the absence of significant differences in certain metrics suggests that the transition from GPT-3.5 to GPT-4 primarily affected the model's engagement with the shared lexicon, but not the complexity or length of the shared expressions themselves (ENTR, L, LMAX).

Another key finding is the symmetry observed between GPT-4 and human participants across all speaker-dependent metrics. This symmetry, evident not only in the shared lexicon metrics, but also in the self-repetition lexicon metrics, indicates a more balanced and reciprocal interaction with human partners. This contrasts with the asymmetry observed with GPT-3.5, which is evident in 7 out of the 9 metrics. GPT-3.5 tended to dominate the conversation (Tokens$_S$), yet was less efficient with its language (VO$_S$), initiated fewer shared expressions (IE$_S$), and repeated itself more often with more complex and longer self-expressions (SER$_S$, SENTR$_S$, SL$_S$, SLMAX$_S$). These behaviors might negatively influence its lexical alignment with humans. As Duplessis et al. noted in their study [1], human-human corpora showcased symmetry in all speaker-dependent metrics, while human-agent corpora exhibited asymmetry in many metrics. Therefore, the symmetry exhibited by GPT-4 suggests a more human-like lexical alignment strategy, potentially contributing to its perceived naturalness and effectiveness in conversations.

The results also revealed the influence of the conversation topic on lexical alignment patterns. Collaborative storytelling conversations exhibited higher values on several metrics, including ER, ENTR, and L, compared to casual conversations, suggesting a more dynamic and varied use of shared expressions. Conversely, casual conversations exhibited a higher value in the overlap of vocabulary, as measured by the VO metric, suggesting a more consistent use of vocabulary. This finding adds nuance to the claim in a previous study [4] that task-oriented conversations tend to involve more convergence than non-task-oriented conversations. It shows that the nature of convergence can vary depending on the topic and the specific metrics examined.

6.2 Limitations and future work

Despite the findings of this study, several limitations should be acknowledged. First, the relatively small size of the corpus used in this study may limit the statistical power and the generalizability of the findings. The limited sample size also makes it difficult to account for individual differences among human participants, such as their language proficiency or personal conversation style, which could potentially influence the patterns of lexical alignment. This limitation poses a risk of inaccuracy in the results; future studies, therefore, could consider involving more participants to form a larger corpus for quantitative analysis.

Second, the interpretations in this study are based solely on quantitative metrics, without the support of qualitative analysis. While these metrics provide valuable insights into the patterns of lexical alignment, they may not fully capture the nuances and complexities of human-GPT interactions. Moreover, the interpretations of such metrics should also be taken with caution, as they are merely possible explanations for the observed results, and the actual mechanisms behind these behaviors in the GPT models are complex and not fully understood. A more comprehensive understanding of lexical alignment could be achieved by incorporating qualitative analyses in future studies, and more research would be needed to confirm these interpretations and explore other potential factors that might contribute to the observed differences in lexical alignment between GPT-3.5 and GPT-4.

Lastly, the study focused on two specific conversation topics. While these topics provide a good starting point, they do not represent the full range of topics that GPT models can engage in. The patterns of lexical alignment may vary significantly across different topics, and future research could benefit from exploring a wider range of conversation topics, both task-oriented and non-task-oriented.

7 CONCLUSION

In conclusion, this study contributes to our understanding of the lexical alignment behaviors of GPT-3.5 and GPT-4 in conversations. The results highlight the significant improvements made in GPT-4, which exhibits a better ability to foster lexical alignment and a more balanced interaction with human participants. The influence of the conversation topic on lexical alignment was also evident, adding nuance to the understanding of how different contexts can influence interaction patterns. These findings underscore the importance of considering both the capabilities of language models and conversation topic in the study of language models' conversational behavior, and the insights gained from this research could inform the development of more engaging and effective conversational agents. This study thus provides a reference point for future research in this field.

REFERENCES

[1] Guillaume Dubuisson Duplessis, Caroline Langlet, Chloé Clavel, and Frédéric Landragin. 2021. Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Lang Resources & Evaluation* 55, 2 (June 2021), 353–388. DOI:https://doi.org/10.1007/s10579-021-09532-w

[2] Heather Friedberg, Diane Litman, and Susannah BF Paletz. 2012. Lexical entrainment and success in student engineering groups. In *2012 ieee spoken language technology workshop (slt)*, IEEE, 404–409.

[3] Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2017. Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *ICIS*.

[4] Patrick GT Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS one* 9, 6 (2014), e98598.

[5] Srinivasan Janarthanam and Oliver Lemon. 2009. A Wizard-of-Oz environment to study referring expression generation in a situated spoken dialogue task. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 94–97.

[6] Anis Koubaa. 2023. GPT-4 vs. GPT-3.5: A concise showdown. (2023).

[7] Theodora Koulouri, Stanislao Lauria, and Robert D. Macredie. 2016. Do (and Say) as I Say: Linguistic Adaptation in Human–Computer Dialogs. *Human–Computer Interaction* 31, 1 (January 2016), 59–95. DOI:https://doi.org/10.1080/07370024.2014.934180

[8] Vivien Kühne, Astrid Marieke Rosenthal-von der Pütten, and Nicole C. Krämer. 2013. Using linguistic alignment to enhance learning experience with pedagogical agents: the special case of dialect. In *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013. Proceedings 13*, Springer, 149–158.

[9] Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, and Annie YS Lau. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.

[10] José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language* 31, 1 (2015), 87–112.

[11] Kate Loveys, Catherine Hiko, Mark Sagar, Xueyuan Zhang, and Elizabeth Broadbent. 2022. "I felt her company": A qualitative study on factors affecting closeness and emotional support seeking with an embodied conversational agent. *International Journal of Human-Computer Studies* 160, (2022), 102771.

[12] OpenAI. 2023. GPT-4 Technical Report. (2023). DOI:https://doi.org/10.48550/ARXIV.2303.08774

[13] Diana Pérez-Marín. 2021. A review of the practical applications of pedagogic conversational agents to be used in school and university classrooms. *Digital* 1, 1 (2021), 18–33.

[14] Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 02 (April 2004). DOI:https://doi.org/10.1017/S0140525X04000056

[15] Martin J. Pickering and Simon Garrod. 2006. Alignment as the Basis for Successful Communication. *Research Language Computation* 4, 2–3 (October 2006), 203–228. DOI:https://doi.org/10.1007/s11168-006-9004-0

[16] Tanmay Sinha and Justine Cassell. 2015. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In *Proceedings of the 1st workshop on modeling INTERPERsonal SynchrONy And influence*, 13–20.

[17] Laura Spillner and Nina Wenig. 2021. Talk to Me on My Level – Linguistic Alignment for Chatbots. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, ACM, Toulouse & Virtual France, 1–12. DOI:https://doi.org/10.1145/3447526.3472050