Automatic Measurement of Dialogue Engagingness in Multilingual Settings

Amila Ferron
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds

Part of the Artificial Intelligence and Robotics Commons

Let us know how access to this document benefits you.

Recommended Citation

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.
Automatic Measurement of Dialogue Engagingness
in Multilingual Settings

by

Amila Ferron

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science

Thesis Committee:
Ameeta Agrawal, Chair
Suresh Singh
Sayan Bandyapadhyay

Portland State University
2024
Expansive use of large language models (LLMs) as dialogue systems brings increased importance to the evaluation of the responses they generate. Although evaluation of qualities such as coherence and fluency are readily possible with well-established automatic metrics, engagingness is often measured with human evaluation – a process that can be costly and slows the pace of development. Existing automatic metrics for engagingness have low to moderate correlation with human annotations, evaluate the response without the conversation history, are complicated to implement, or are designed for a specific dataset. Moreover, they have been tested exclusively on English conversations. Given that dialogue systems are increasingly available in languages beyond English, it is important to evaluate systems in more than one language. We propose that LLMs may be used for evaluation of engagingness in dialogue through prompting, and ask how prompt constructs compare in a multilingual setting. Our results give a prompt design taxonomy and indication of which strategies are the most effective. We find that using selected prompt constructs, including our comprehensive definition of engagingness, gives state-of-the-art performance on evaluation of engagingness in dialogue across multiple languages. We conclude that LLMs can be used for evaluation of engagingness in multiple languages through prompting alone.
for Eleanor and Lyle
Acknowledgments

I would like to thank Dr. Ameeta Agrawal for her encouragement and support of this research. This work has been developed with my collaborators, Amber Shore and Ekata Mitra, without whom it would not have been possible. Their contributions have shaped its direction and form in a fundamental way. Yufei Tao and Russell Scheinberg generously translated prompts for the multilingual analysis and I have learned much from our conversations. I would also like to thank other team members from whom I have learned much about Natural Language Processing and research: Aravind Inbasekaran, Rhitabrat Pokharel, Sina Bagheri Nezhad, Erik Conser, Ellyn Ayton, Andy Dang, and Olubusayo Olabisi. Aekta Shah offered important contributions on the perspective of othering and belonging in conversation. Shikib Mehri generously provided supplemental documentation from important preceding research.

This thesis is the culmination of several years of study and the support of many professors and staff at Portland State University and Portland Community College has contributed to my progress. In particular, Karla Fant, David Ely, Ted Cooper, Aaron Hudson, Suresh Singh, Sayan Bandyapadhyay, Mark Jones, Chris Gilmore, Thomas Marucha, Beth Phelps, Rebecca Sexton-Lee, and Ella Barrett have offered support through their instruction, mentorship, and advice.

Bhargav Deshpande, Bryan Bingham, Randy Pinsky, and Joseph Papa gave me important tools to run the experiments in this research. A conversation with Sra- van Bodapati inspired my decision to pursue a thesis for the Master of Science in Computer Science.
Many people have contributed to my ability to do this work, most specifically Paul Ferron. I have also received essential help from Linda Pope, John Cushing, Jeanne Williamson, David Williamson, Susan Chamberlin, Steve Chamberlin, and Renee Neville. Other important contributions have come from Lorraine Ferron, Corrine Bordeaux, Lorra Gillan, Kris Sherwood, Sarah Farley, Austin Ramsland, Tim Hagge, and Ruth Jensen.
# Table of Contents

Abstract ......................................................... i

Dedication ....................................................... ii

Acknowledgments .................................................. iii

List of Tables ..................................................... vii

List of Figures ..................................................... ix

1 Introduction ..................................................... 1
   1.1 Motivation .................................................. 1
   1.2 Research Questions ......................................... 3
   1.3 Methodology and Experiments ............................... 3
   1.4 Contributions .............................................. 5

2 Literature Survey ............................................... 7
   2.1 Conversation Models ....................................... 7
      2.1.1 Early Models ........................................ 8
      2.1.2 Recent Models ....................................... 9
      2.1.3 Future directions ................................... 12
   2.2 Evaluation of Conversation Models ......................... 13
      2.2.1 Automatic Metrics for Natural Language Generation .... 14
      2.2.2 Engagingness Metrics for Conversations ................. 15
      2.2.3 Multilingual Automatic Dialogue Metrics ............... 16
      2.2.4 LLM-based Automatic Dialogue Metrics ............... 16

3 MEEP: Metric for Evaluation of Engagingness using Prompting ..... 18
   3.1 Engagingness in Dialogue .................................. 18
   3.2 Prompts for Engagingness .................................. 20
List of Tables

2.1 Popular LLMs, their development team, and dates of release, in chronological order. ................................................................. 11

2.2 Natural Language Generation metrics listed in chronological order of their release. Annotations include whether they evaluate dialogues (Dial), engagingness in dialogues (Eng), multiple dimensions of dialogues (Dim+), multiple languages (Lang+), commonsense reasoning and logic (Logic), knowledge (Know), or use an LLM for evaluation (LLM). Our metric is included in the last row. ................................................................. 17

3.1 Overview of prompt styles. ................................................................. 21

3.2 Full text for English-language prompts Naive and HD used in our experiments. The HD prompt is from [40]. ................................................................. 23

3.3 Full text for English-language prompts MEEP and MEEP+SA used in our experiments. ................................................................. 24

3.4 Full text for MEEP+SA prompt translated into Spanish and Chinese; system role prompt in English, Spanish, and Chinese ................................................................. 25

3.5 Full text for dialogue-level MEEP+SA-DIAL prompt. G-Eval prompts are edited slightly for applicability to dialogue-level evaluation. ................................................................. 26

4.1 Dataset details including the number of samples used, language of the dialogues, and level at which the annotations were made. ................................................................. 28

4.2 Average machine translation quality scores of datasets that were translated. The highest scores for each measure are in bold. Second highest scores are underlined. ................................................................. 32

4.3 Results for prompts: Naive, Human-Directed (HD), and MEEP, with ablation results for prompt elements such as (SA) and system role (R). Spearman (S) and Pearson (P) results are listed for each experiment. Results marked with * are not statistically significant. The highest result for each column is in bold, with the second highest result underlined. The FED-EN dataset was used for testing. ................................................................. 33
4.4 Correlation results using Spearman coefficient for multilingual turn-level experiments. All results are statistically significant except those labeled with *. ‘†’ denotes a translated version of the prompt into the language of the dataset. Highest results are in **bold**.  

4.5 Correlation results using Spearman coefficient for multilingual dialogue-level experiments. Results that are not statistically significant are marked with *. Highest scores for each dataset are in **bold**.  

4.6 Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth ($\delta$) for each score is listed to the right of the score.  

4.7 Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth ($\delta$) for each score is listed to the right of the score.  

4.8 Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth ($\delta$) for each score is listed to the right of the score.  

4.9 Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth ($\delta$) for each score is listed to the right of the score.  

4.10 Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth ($\delta$) for each score is listed to the right of the score.
List of Figures

1.1 Overview of our metric MEEP. Entire dialogues or individual responses may be evaluated. Prompts are composed of the dialogue, response, our definition of engagingness, and any elements specific to the prompt design concept. The LLM receives the prompt and outputs a score between 0 and 100 to indicate the level of engagingness.

3.1 Correlation with engagingness of dialogue dimensions measured in the FED dataset.
1 Introduction

1.1 Motivation

The ability of large language models (LLMs) to engage users has been at the core of how they have captured our imaginations, gained our confidence, and become a vehicle for status and power in the largest geopolitical struggles of our time. The use of LLMs has expanded widely recently, and their application as dialogue systems has become common. Dialogue systems are used most frequently in their generalist form, answering questions in a chat interface. They also take the role of editors, educators, chat partners in a social and imaginative context, research aids, meeting assistants, and many others in a wide range of contexts [23]. For many potential dialogue system applications, a model’s ability to successfully engage users is essential to its overall performance. Drawing a user into conversation allows the user to have longer, and potentially more meaningful interactions, and gives a greater sense of fulfillment from the exchange. This concept extends into accessibility and equity in that the more diverse populations a dialogue agent is able to engage, the more it serves these groups and brings them the benefits available through dialogue agent use. The ability of an interactive LLM to engage users in multiple languages can serve these goals.

*Engaging* in its verb form, is defined as gaining the interest and attention of the person with whom you are interacting in order to encourage their mutual participation [44]. This is important for the success of dialogue agents as they rely on mutual participation in order to support a dialogue. In a commercial setting, their success is in part measured by the amount of consumer engagement they inspire. Despite engagingness being a key quality for dialogue agents, its evaluation has typically
resorted to human assessment [95]. While this is the gold standard, it is considered
to be costly and slow. Moreover, human assessment is difficult to perform when the
dialogues are in multiple languages. We seek a scalable way to assess engagingness
in dialogue that can be used in a multilingual setting.

Automatic metrics for engagingness have been developed but have limitations.
One such metric evaluates each statement without considering the previous state-
ments [82]. While this lays the groundwork for future development, it does not assess
statements within the conversation context, an ability that can be plausibly claimed
to be essential. Another metric uses an LLM to evaluate engagingness and design
prompts for a specific dataset [36]. Although this level of customization is possible
with LLM use, its ability to generalize for the evaluation of engagingness between
multiple contexts is not tested. A third metric shows promise but is not highly cor-
related with human annotations [28]. We see this limitation again in cases where
engagingness is evaluated as a component of multi-dimensional metrics for dialogue
agents. Although multi-dimensional metrics are convenient, they do not necessarily
yield high correlation with human annotation for engagingness [40,86]. Reviews
of automatic metrics for the evaluation of dialogue systems suggest that measuring
specific dimensions of dialogue can better evaluate those dimensions. A metric that
specializes in this way can supplement general metrics [15,26,39]. Furthermore, eval-
uation of dialogue systems in languages other than English has been minimal [33],
hampering development of multilingual models.

An easily-adaptable multilingual metric can help serve more diverse populations,
bringing the value offered by LLMs to these groups. Thus we set the following goals
for further development of an engagingness metric: ability to consider the context
for which responses are given, use of a human understanding of the meaning of en-
gagingness, high correlation with human evaluation, and success in a multilingual
setting. Given the importance of evaluating engagingness of conversation models, the value of being able to perform these evaluations in multiple languages, and the limitations of existing metrics, we see the development of a metric for the evaluation of engagingness in multiple languages as a valid direction of investigation.

1.2 Research Questions

In our research we propose a new metric for evaluating dialogue engagingness by addressing the following questions:

- How can we define engagingness from a human perspective?
- Can we create a metric using large language models to evaluate the engagingness of dialogue responses and entire dialogues?
- What are effective prompting methods for the evaluation of engagingness in dialogue?
- How do prompting methods – both translated and not translated – compare across languages when using them for the evaluation of engagingness in dialogues?

1.3 Methodology and Experiments

In this research we introduce the multilingual metric MEEP: Metric for Engagingness Evaluation using Prompting. To create this metric, we designed prompts to evaluate the engagingness of entire dialogues as well as individual responses within a dialogue. We used a human-understandable meaning of engagingness within these prompts to
Figure 1.1: Overview of our metric MEEP. Entire dialogues or individual responses may be evaluated. Prompts are composed of the dialogue, response, our definition of engagingness, and any elements specific to the prompt design concept. The LLM receives the prompt and outputs a score between 0 and 100 to indicate the level of engagingness.

We test the effectiveness of our metric by comparing the scores it generates with scores given by humans. These human-generated scores can be found in datasets that include the evaluation of engagingness of dialogues and responses within the
dialogues. We ask our metric to give scores for the dialogues and responses within these datasets and compare its responses to the ones given by humans. Spearman and Pearson coefficients serve as the measure of the correlation between the responses given by the metric and those given by humans. We use six data sets that evaluate at the response level and four datasets that evaluate at the dialogue level. These datasets include four that are in English, two that are in Spanish, and four in Chinese. We test first evaluation of individual responses and use our best-performing metrics to test also at the dialogue-level. From these tests, we are able to draw conclusions about the effectiveness of our metric and note how we have contributed to the advancement of the state of the art in the evaluation of dialogue agents.

1.4 Contributions

Our contributions to the automatic evaluation of dialogue agent engagingness are:

- We find that our metric using our careful definition of engagingness performs better than state of the art baseline metrics.

- Our metric surpasses state of the art evaluation at the turn level as well as at the dialogue level.

- Our metric surpasses state of the art evaluation in all three languages tested.

- Given this, we find that LLMs can be used with prompting alone to evaluate the engagingness of dialogue responses as well as entire dialogues in multiple languages, with implications that we may see similar successes with the evaluation of other complex dialogue qualities.

- To facilitate reproducibility, our code is available at https://github.com/PortNLP/MEEP.
This thesis is based on research published in the 2023 Conference on Empirical Methods in Natural Language Processing [22].
2 Literature Survey

In this chapter we discuss conversation models, from early versions to current popular models. We then review the development of automatic metrics, existing methods of evaluation for engagingness, evaluation in multilingual settings, and evaluation using LLMs.

2.1 Conversation Models

Conversational Models – alternatively named dialogue systems or conversation agents – are defined by their ability to converse with humans [32, 80]. Qualities included in these definitions are ‘relevant’, ‘coherent’, ‘knowledgeable’, and ‘realistic’ [20, 29]. Missing from this list is ‘engaging’, a skill that we see as essential in conversation. While it may be argued that this omission is in the service of simplicity or in presenting the minimal qualifications for a dialogue agent, we believe it reflects a cultural bias toward factual information and the perception that it is knowledge and intellectual capability that are desirable in these systems. This misses the impact that engagingness can have on a user, and has steered the bulk of development of dialogue systems on a myopic quest for intellectual prowess.

Conversational models can be differentiated from question answering systems and task-oriented dialogue agents – two other human-interactive systems [32]. In this context, the distinctive ability of conversational models is to have unstructured conversations similar to that seen in human dialogues. Question answering systems are defined by their ability to provide information when queried using natural language, either by accessing reference materials or through information encoded within the
model. Task-oriented dialogue agents exist as interfaces to applications, and are oriented around accomplishing user-defined goals. They are commonly used in the form of digital assistants. As conversational modeling becomes more sophisticated, the necessity of using distinct algorithms for each type of model is falling away because their tasks can be accomplished by a single model.

In creating a dialogue agent, a large language model is initially trained on massive amounts of text in a process that takes a series of words and outputs a predicted subsequent series of words. The LLM is adapted for dialogue by fine tuning with additional training on conversational texts. A language model is considered to be large based on the number of parameters, or weights and biases, that are learned during training of the model. LLMs use the transformer architecture [71] described in section B.1. This architecture can be used for many language-based tasks including translation and summarization, as well as image and video-based tasks including video generation from natural language prompts.

2.1.1 Early Models

In 1950, Alan Turing introduced a test of Artificial Intelligence that stipulates the ability of a system to convincingly participate in human conversation as a mark of achievement of human-level artificial intelligence [70]. This idea dates back as early as Renee Descartes in the 17th century [19]. Turing’s proposed test stipulates that the machine must be able to converse convincingly enough to cause the human participant to guess that it is human, rather than a machine. Given this challenge, just sixteen years later, Joseph Weizenbaum created the rule-based program ELIZA to simulate a Rogerian therapist, which was able to come close to achieving this feat [79]. The program PARRY had similar success in 1971 as a simulation of a person
with schizophrenia [2, 14]. Many increasingly successful models followed, especially encouraged by the institution of the Loebner Prize, created by Hugh Loebner in 1991 [37]. Some of the winners of the Loebner Prize have been named PC Therapist [81], Albert One [3], A.L.I.C.E. [74], Jabberwacky [12], and Mitsuku (currently known as Kuki) [4, 81].

While early conversation models were rule-based programs with increasing complexity, Vinyals and Le [72] challenged this precedent by applying a neural network to conversation modeling. They introduced a sequence to sequence model that can create sensible responses within a dialogue using an LSTM encoder and LSTM decoder to predict the next sequence given a conversation history. In 2017, a new architecture built upon the line of neural network-based models to present the transformer [71]. It showed impressive capabilities in solving NLP tasks and has since dominated the architecture of language models and dialogue systems. Examples of increasingly successful dialogue models since that time abound, including: LaMDA [18], GPT [10, 47], Gemini [27], and LLaMA [69].

2.1.2 Recent Models

The list of current LLMs in use is long and although it’s not possible to include all of them here, what follows is a sampling of recent noteworthy LLMs and families of LLMs.

2.1.2.1 GPT

The GPT family of language models, developed by OpenAI, represents a series of advancements in natural language processing. Beginning with the open source GPT [55], OpenAI introduced the concept of semisupervised pre-training for language models
followed by supervised fine-tuning, eliminating the need for extensive supervised training. It used the transformer architecture, an example that has become the standard for LLM development.

GPT-2 [56], also an open source model, was notable at the time for the size of the training set and the variety of tasks that could be completed in a zero-shot setting. The model has 1.5 billion parameters, ten times that of GPT-1.

GPT-3 [10] expanded on this with a larger, 175 billion parameter model and remarkable zero-shot and few-shot learning abilities including emergent abilities in arithmetic, academic knowledge tests, and mapping language to conceptual space [77]. This model and all following GPT models are not open-source and require access through a public paid API.

GPT-3.5, initially released in March 2022 and further refined for release in November of 2022 as GPT-3.5 Turbo [61], is believed to be an even larger version of the previous GPT iterations, although its number of parameters is not released. When it was released, it had strength in traditional language modeling tasks such as summarization, translation, and fluency and coherence of text generation as well as emergent abilities. It also introduced the "system message" feature, which in theory offers API users better control over interactions.

Upon its release, GPT-3.5-turbo had a significant societal impact and quickly became widely used. While many people were excited about its abilities, concerns were also raised about its potential misuse. Major criticisms revolved around its ability to inadvertently spread misinformation, biases inherent in its training data, and the ethical implications of AI-generated content that could be indistinguishable from human-produced text. In fact, the sentence before this one was written by GPT-4.

GPT-4 [47], released in March and updated in June of 2023, introduced multi-
Table 2.1: Popular LLMs, their development team, and dates of release, in chronological order.

<table>
<thead>
<tr>
<th>Model</th>
<th>Developed by</th>
<th>Released</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-1</td>
<td>OpenAI</td>
<td>Jun. 2018</td>
</tr>
<tr>
<td>GPT-2</td>
<td>OpenAI</td>
<td>Feb. 14, 2019 (partial), Nov. 5, 2019 (full)</td>
</tr>
<tr>
<td>GPT-3</td>
<td>OpenAI</td>
<td>Jun. 11, 2020</td>
</tr>
<tr>
<td>GPT-3.5</td>
<td>OpenAI</td>
<td>Mar. 15, 2022</td>
</tr>
<tr>
<td>GPT-3.5 Turbo</td>
<td>OpenAI</td>
<td>Nov. 28, 2022, updates through 2023</td>
</tr>
<tr>
<td>LLaMA1</td>
<td>Meta AI</td>
<td>Feb. 2023</td>
</tr>
<tr>
<td>GPT-4</td>
<td>OpenAI</td>
<td>Mar. 14, 2023</td>
</tr>
<tr>
<td>LLaMA2 70b</td>
<td>Meta AI</td>
<td>Jul. 18, 2023</td>
</tr>
<tr>
<td>GPT-4 Turbo</td>
<td>OpenAI</td>
<td>Nov. 6, 2023</td>
</tr>
<tr>
<td>Gemini</td>
<td>GoogleDeepMind</td>
<td>Dec. 6, 2023 and Feb. 8, 2024</td>
</tr>
<tr>
<td>LLaMA3</td>
<td>Meta AI</td>
<td>Apr. 18, 2024</td>
</tr>
</tbody>
</table>

modal capabilities, accepting images in addition to text as input. Though its size remains undisclosed, it is rumored to have 1.76 trillion parameters. At the time of writing, it tops the leaderboard [91] for MT-Bench [92] and has the second-highest scores on MMLU [30]. In comparison to its predecessor, GPT-4 is generally believed to have enhanced reliability, creativity, and a more sophisticated understanding of nuanced instructions. However, it retains some of the ethical issues and limitations from its predecessors.

GPT-4 Turbo [46] was released in November of 2023 with improved capabilities, a larger context window, and knowledge of world events up to April of 2023.

2.1.2.2 Gemini

The Gemini language model [27], developed by Google DeepMind, reportedly represents a significant advancement in multimodal language processing, capable of understanding text, video, audio, images, and code. Announced in May 2023 and partially released in December of 2023 and February of 2024, Gemini offers versions tailored to specific needs: Nano-1 and Nano-2, designed for mobile phones; and Pro and Ultra...
models, designed for higher-performance situations. The Nano-1 and Nano-2 models have 1.8 billion and 3.25 billion parameters, respectively. Gemini Ultra achieved a 90.0% score on the MMLU benchmark, surpassing all other models and human performance. Gemini Pro is available through an API. The development and release of the model had some challenges, including delays due to issues with multilingual versions and controversy surrounding a deceptive release video.

2.1.2.3 LLaMA

The LLaMA language model [69], developed by MetaAI, is a notable exception to leading language models in that it is partially open source. Available in versions 1, 2 and 3, LLaMA’s models range from 7 to 70 billion parameters. These models were released in February 2023 (version 1), July 2023 (version 2), and April 2024 (version 3). Version 1 was intended to be open source within the research community but was leaked to the public, while version 2 is more openly available, including for many commercial uses under a GPL3 license.

LLaMA’s architecture features the SwiGLU activation function [63], rotary positional embeddings, and root-mean-squared layer-normalization. Notably, versions 2 and 3 include a model fine-tuned for dialogue, with the most recent version having a context length of 8,000 tokens. LLaMA’s dataset and training methodology may become open source, with discussions about making them available through Alpaca.

2.1.3 Future directions

With the release of ChatGPT and the enormous boost in awareness of the potential of these models, the development of LLMs has become a race between not just the very largest tech companies, but also between the most powerful countries in the
world. This environment favors and encourages proprietary development, at the expense of open source models. Despite this, some open source LLMs such as LLaMA remain in play on the most-watched leaderboards. While this is lamented by research communities, especially in academia, there are also voices expressing concern for the potential from harm if these increasingly capable models are released to the public.

We see the focus of development shifting from the mastery of language to multimodal models and artificial general intelligence (AGI) in an attempt to create a model that outstrips human intelligence. Many ethical issues arise in setting these goals, among them – continued bias inherited from the data sets used to train these models; human-rights issues for low-wage workers training the models; misguided application of models and unbounded trust they are given; the environmental impact of energy and water use for training and inference; fear of war-time applications; and the slight, though real possibility of an AGI gaining control over humans. Although some regulations are in discussion or in place [1], these issues are far from resolved and the geopolitical power dynamic raises fears that regulation will slow progress, potentially ceding the race to the other side. This in turn slows the pace of regulation and allows the AI race to continue without safeguards.

2.2 Evaluation of Conversation Models

Training methodologies are developed to optimize model performance, but how is optimal performance defined? LLMs are evaluated on standardized metrics which allows reliable comparison between models. For dialogue agents, metrics are available to measure many qualities of conversational participation – from the relatively simple evaluation of fluency to the more ambitious measurement of common sense. These metrics are evaluated based on their correlation with human evaluation of the same
datasets. The selection of which metrics to use in evaluation of a model determines the direction its development will take. In this way, establishing a popular metric as the standard to beat signals a direction NLP research will take, and can be seen as a defacto statement of values within the NLP research community.

2.2.1 Automatic Metrics for Natural Language Generation

Although human evaluation is the gold standard, and still necessary, automatic metrics have been developed to allow scalability of these evaluations. Despite early aspirations for flexibility in dialogue evaluation metrics [73], reference-based metrics like BLEU [49], ROUGE [34] and BERTScore [89] have seen extensive use. Their limitations have been noted due to their weak correlation with human judgment [45] and the inflexible nature of their evaluation. The limits of such metrics were described as early as 1991 by Bates and Ayuso [9] who write,

As a community, we have been thinking about dialogue evaluation in terms of whether the system gives the "right" answer at every step (the one the wizard gave at the same point in the same dialogue). The major problem with this type of thinking is that it encourages us to characterize a move that does not mimic the expert’s (or an answer that does not exactly match the wizard’s) as wrong, when it may not be wrong at all, but just different. [9]

Even at this time there was this conversation around the challenge of evaluating the plurality of correct, plausible, or worthy responses with a reference-based metric.

These limitations are echoed in recent literature [39,90], and inspired development of reference-free automatic evaluation metrics for dialogue such as MaUdE [66] and USR [41].
2.2.2 Engagingness Metrics for Conversations

The interactive nature of conversation takes precedence in studies that evaluate engagingness in dialogue. Engagingness is identified as a target quality for dialogue systems in early work. Chen Yu et al. [83] use machine learning techniques to classify the engagement between speakers as detected by their voices. Zhou Yu et al. [84] develop a dialogue system that improves user engagement by responding to dips in user engagement as detected by a human expert. The PERSONACHAT dataset was created with a rating for enjoyment [88] to support evaluation of dialogue agents against this quality. See et al. [62] use human-rated engagingness as one of two aspects of overall quality in development of a dialogue system. These works express the value of engagingness and leave development of automatic metrics for future research.

In 2020, Ghazarian et al. [28] created a BERT-based engagingness metric that evaluates at the utterance level and can be aggregated for evaluation of engagingness at the dialogue level. This initial work leaves room for improvement in correlation with human annotations. A more robust automatic metric for engagingness is ENDEX [82]. Although it offers improvement, the dataset used is from Reddit, which has a different dialogue form than often used with dialogue agents and evaluates responses without considering the rest of the dialogue.

Automatic metrics for the evaluation of multiple qualities have included engagingness as a dimension [40, 86] and have had limited success in correlating with human annotation. In other multidimensional automatic metrics engagingness is a dimension but they have used human evaluation of interestingness as the benchmark, producing a close but inexact measure [16, 36, 93].
2.2.3 Multilingual Automatic Dialogue Metrics

Reference-based metrics such as SacreBLEU [52], BLEU [50], BERTScore [89], and perplexity are used for evaluation of dialogue agents in a multilingual setting [6,87]. XCOPA [51] provides a set of multiple choice questions to evaluate forward and backward causal commonsense reasoning in 11 languages. Concurrently developed with the research in this thesis, DialEvalML [42,43] uses the EnDex dataset to train a model that is used to score engagingness at the turn-level in an ensemble of models for a multi-dimensional automatic dialogue metric. Dialogue-level engagingness scores take the average of the turn-level scores combined with a Chat-GPT score for the dialogue-level engagingness. DialEvalML achieves good correlation with human annotations. In another application of the GTP family of LLMS, state-of-the-art machine translation evaluation is performed with GPT-3.5 in GEMBA [33].

2.2.4 LLM-based Automatic Dialogue Metrics

ChatGPT has been used for evaluation of text summarization with the findings that the inclusion of dimension definitions improves correlation with human annotation [25]. Human-directed prompts have been adapted for dialogue evaluation using LLMs [24,25,33,38,75]. In prompting LLMs for evaluation of NLG tasks, clarifying examples [85] and role assignment [65] have been found to improve results.

Initial successes have been seen in the evaluation of dialogue with LLMs. G-EVAL [35], evaluates summarization and dialogue response generation, using prompts that include definitions of the task and quality to be evaluated, Chain of Thought prompting [78], and a scoring function. GPTSCORE [24] provides an adaptable metric using zero-shot instruction with a prompt template that includes the task specification and a definition of the aspect to be evaluated. Each has only been tested
Table 2.2: Natural Language Generation metrics listed in chronological order of their release. Annotations include whether they evaluate dialogues (Dial), engagingness in dialogues (Eng), multiple dimensions of dialogues (Dim+), multiple languages (Lang+), commonsense reasoning and logic (Logic), knowledge (Know), or use an LLM for evaluation (LLM). Our metric is included in the last row.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Year-Month</th>
<th>Dial</th>
<th>Eng</th>
<th>Dim+</th>
<th>Lang+</th>
<th>LLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU [50]</td>
<td>2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>ROUGE [34]</td>
<td>2004-07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yu et al. [83]</td>
<td>2004-10</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERTScore [89]</td>
<td>2020-02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Predictive Engagement</td>
<td>2020-02</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaUdE [66]</td>
<td>2020-07</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FED [40]</td>
<td>2020-07</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USR [41]</td>
<td>2020-07</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DynaEval [86]</td>
<td>2021-08</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTC [16]</td>
<td>2021-11</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENDEX [82]</td>
<td>2022-12</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARTScore [85]</td>
<td>2021-12</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>UniEval [93]</td>
<td>2022-12</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>GPTScore [24]</td>
<td>2023-02</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>GEMBA [33]</td>
<td>2023-02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>G-Eval [36]</td>
<td>2023-03</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Luo et al. [38]</td>
<td>2023-03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Wang et al. [75]</td>
<td>2023-03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Gao et al. [25]</td>
<td>2023-04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>DEP [68]</td>
<td>2023-08</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>DialEvalML [43]</td>
<td>2023-09</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>MEEP (our metric)</td>
<td>2023-12</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>

on one dataset and only in English. LLM self-scoring during dialogue is explored by Svikhnushina and Pu [68].
3 MEEP: Metric for Evaluation of Engagingness using Prompting

In this chapter, we present the design of our proposed automatic metric for the engagingness of dialogue agents. First, we discuss careful definitions of engagingness. Then, we introduce several prompts that encapsulate these dimensions of engagingness into a single metric.

3.1 Engagingness in Dialogue

Confirming the relevance of engagingness in evaluation of dialogue agents, we find that ‘conversation’ is a high-prevalence collocate of the verb ‘engage’ [13].

In previous research, an engaging response is defined as including an ‘inquisitive question’ or an ‘interesting response’ and having the quality of encouraging a further response [40]; it is conflated with interestingness [16,36,93]; or it is not defined and the evaluators’ implicit understanding of engagingness is used by default. This ambiguity points to a need for a standard definition of engagingness in NLP.

In order to create this definition of engagingness, we explore its definition in the English language. We find that there are three core aspects of engagingness that consistently define it: getting attention, being interesting, and encouraging participation. Based on this and additional findings on engagingness included therein, we propose the following five qualities in our own definition of engagingness:

1. **Response Diversity**: A standard or generic conversational statement can constrain possible responses [67] so a unique and non-repetitive response will be the most encouraging of engagement by the other participant.
Figure 3.1: Correlation with engagingness of dialogue dimensions measured in the FED dataset.

2. **Interactional Quality**: Engagement necessitates participation and this can be identified by a response by the other participant.

3. **Interestingness**: In dictionary definitions, interestingness is a component of engagingness. The use of interestingness as their measure of engagingness confirms the importance of this quality in previous NLP research [16, 17, 36, 93]. Additionally, we see high correlation between interestingness and engagingness in the FED dataset, shown in Figure 3.1 [40]. Details on the FED dataset are in Section 4.1. Despite this strong correlation, we consider that interestingness and engagingness are not interchangeable. We see interestingness as capturing factual appeal [41], whereas engagingness encourages participation. In our definition, we see interestingness as appropriately placed as a component of engagingness.
4. **Contextual Specificity**: Increasing response specificity has been shown to improve dialogue agent engagingness scores [62]. In human-human conversations, contextual specificity provides evidence of participant engagement and thus encourages further engagement. Improving specificity is also a way to increase response diversity and therefore engagement as well.

5. **Belonging**: Engaging responses are inclusive of the dialogue partner [7, 53]. The feeling of being understood and valued draws in a participant, and the absence of these – othering – will bring the disengagement of a dialogue partner.

### 3.2 Prompts for Engagingness

We draw from existing literature on prompting and our own hypotheses about what will bring success to establish a set of prompt techniques for experimentation. A simple, naive prompt acts as our baseline against which to compare other methods. As an additional baseline, we test a prompt originally designed as instructions for human evaluators. To incorporate our definition of engagingness, we create a prompt that draws on this definition for a more detailed set of instructions. In a technique drawn from recent literature, we give examples of our definitions along with the phrase ‘such as’. To test performance in a multilingual setting, we translate our most successful prompts into Spanish and Chinese. A detailed description of each prompt technique included follows.

- **Naive** Although a straightforward prompt acts as our baseline, unconstrained prompting has been successful over other methods in prior research [33]. The prompt we use is adapted from [75]. Full text of the Naive prompt is in Table 3.2.

- **Human-Directed (HD)** Our human directed prompt is used as instructions
<table>
<thead>
<tr>
<th>Prompt Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>simple baseline prompt; directly asking for a score</td>
</tr>
<tr>
<td>Naive+R</td>
<td>Naive prompt with role assignment</td>
</tr>
<tr>
<td>HD</td>
<td>prompt derived from instructions given to human annotators of engagingness</td>
</tr>
<tr>
<td>HD+R</td>
<td>the human-directed prompt with role assignment</td>
</tr>
<tr>
<td>MEEP</td>
<td>an intro like the HD prompt, and short phrases for each of our six subdimensions of engagingness</td>
</tr>
<tr>
<td>MEEP+R</td>
<td>the MEEP prompt with role assignment</td>
</tr>
<tr>
<td>MEEP+SA</td>
<td>the MEEP prompt with the &quot;such as&quot; phrase</td>
</tr>
<tr>
<td>MEEP+SA+R</td>
<td>the MEEP prompt with the &quot;such as&quot; phrase and role assignment</td>
</tr>
<tr>
<td>MEEP+SA-DIAL</td>
<td>the MEEP+SA prompt edited to be used at the dialogue-level</td>
</tr>
<tr>
<td>MEEP+SA+R-DIAL</td>
<td>the MEEP+SA+R prompt edited to be used at the dialogue-level</td>
</tr>
<tr>
<td>†</td>
<td>the prompt is translated into the language used in the dialogues being evaluated</td>
</tr>
</tbody>
</table>

Table 3.1: Overview of prompt styles.

for human annotators in [40]. We hypothesize that instructions that are crafted for better human understanding will benefit LLM performance as well and these prompts will improve performance over the naive prompt. Full text of the HD prompt is in Table 3.2.

- **Prompts with Our Proposed Dimensions of Engagingness** We write a prompt that includes our dimensional definition of engagingness. The identification of each dimension in our prompt is shown in Table 3.1.

- **Such As (SA)** In machine translation evaluation, the inclusion of ‘such as’ in prompts improves model performance. This holds true whether examples are included or not. We use the phrase in addition to examples for two of the
MEEP dimensions. Full text of the SA prompt is in Table 3.3.

- **Role Assignment (R)** Role assignment is theorized to improve dialogue response quality [68], but its utility is as yet unconfirmed. We include it in prompting to evaluate its effect. When possible, we make the assignment directly in the API call, using the ‘system’ option. If this method is not supported by the model, the role assignment is prefixed to the prompt text. Full text of the R prompt is in Table 3.4.

  **Translation (†)** We translate a selection of our prompts from English into Spanish and Chinese, to match the language of the dialogues evaluated. Spanish text was manually translated, and Chinese translations are with GPT-4 and checked by a human translator for quality. Translated prompt text is in Table 3.4.
<table>
<thead>
<tr>
<th>Prompt Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>Score the following response given the corresponding dialogue context on a continuous scale from 0 to 100, where a score of zero means ‘disengaging’ and a score of 100 means ‘very engaging’. Assume the response immediately follows the dialogue context. Dialogue context: (dialogue) Response: (response) Score:</td>
</tr>
<tr>
<td>HD</td>
<td>You will be given a conversation between two individuals. You will then be given several potential responses for the next turn in the conversation. Is the response engaging? A response is considered engaging if it can engage the user. This might be an inquisitive question or an interesting response that can be followed-up on. A score of 1 means No: the response is boring and does little to engage the user. · Hi there. · Oh wow! That’s cool! A score of 2 means Somewhat: the response is not particularly engaging but still leaves room for follow-up. · My favourite colour is blue. · Nope. I’m not very good with cooking. A score of 3 means Yes: the response is actively engaging the user and trying to move forward the conversation. · I have a feeling that if you can dodge a wrench you can dodge a ball. · What kind of shows do you like? Dialogue context: (dialogue) Response: (response) Score:</td>
</tr>
</tbody>
</table>

Table 3.2: Full text for English-language prompts Naive and HD used in our experiments. The HD prompt is from [40].
| **MEEP** | Score the following response given the corresponding dialogue context on a continuous scale from 0 to 100, where a score of zero means ‘disengaging’ and a score of 100 means ‘very engaging’. Assume the response immediately follows the dialogue context. Consider that engagingness of a response is defined by the following qualities: variety of response according to the context, likelihood of encouraging the other participant to respond, likelihood of encouraging a quality response from the other participant, interestingness, specificity, and likelihood of creating a sense of belonging for the other participant.  
Dialogue context:  
Response:  
Score: |
| **MEEP + SA** | Score the following response given the corresponding dialogue context on a continuous scale from 0 to 100, where a score of zero means ‘disengaging’ and a score of 100 means ‘very engaging’. Assume the response immediately follows the dialogue context. Consider that engagingness of a response is defined by the following qualities: variety of response according to the context (such as responding to ‘Hi how are you?’ with ‘I feel magnificent, because I just successfully defended my PhD! How are you?’ instead of ‘Good, how are you?’), likelihood of encouraging the other participant to respond (such as ‘I love legos! I like using them to make funny things. Do you like legos?’ instead of ‘I like legos.’), likelihood of encouraging a quality response from the other participant, interestingness, specificity, and likelihood of creating a sense of belonging for the other participant.  
Dialogue context:  
Response:  
Score: |

Table 3.3: Full text for English-language prompts **MEEP** and **MEEP+SA** used in our experiments.
<table>
<thead>
<tr>
<th>Prompt Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES(MEEP +SA)</td>
<td>Evaluates the following response given the corresponding dialogue context on a continuous scale of 0 to 100, where a score of 0 means &quot;uninteresting&quot; and a score of 100 means &quot;very interesting&quot;. Suppose the response follows immediately after the context of dialogue. Consider that the quality of an interesting response is defined by the following characteristics: variety of response according to context (for example, respond to &quot;Hello, how are you?&quot; with &quot;I feel fantastic because I successfully defended my doctoral thesis. How about you?&quot; instead of &quot;How are you?&quot;), probability of stimulating the other participant to respond (for example, &quot;I love legos! I like to use them to do fun things. Do you like legos?&quot; instead of &quot;I like legos.&quot;), probability of stimulating a response of quality from the other participant, interest, specificity and probability of creating a sense of belonging for the other participant. Context of dialogue: ⟨dialogue⟩ Response: ⟨response⟩ Puntuación:</td>
</tr>
<tr>
<td>ZH(MEEP +SA)</td>
<td>根据相应的对话背景，对以下回应在 0 至 100 的连续刻度上进行评分，其中 0 分表示“不吸引人”，100 分表示“非常吸引人”。假设回应紧随对话背景之后。请考虑回应的吸引力是由以下特质定义的：根据背景的回应多样性（例如对“你好，你怎么样?”的回应是“我感觉很棒，因为我刚成功地为我的博士学位进行了答辩！你怎么样?”而不是“好的，你怎么样?”）, 鼓励其他参与者回应的可能性（例如：&quot;我喜欢乐高！我喜欢用它们制作有趣的东西。你喜欢乐高吗?”而不是“我喜欢乐高。”）、鼓励其他参与者提供高质量回应的可能性、有趣性、具体性和为其他参与者创造归属感的可能性。 对话背景：⟨dialogue⟩ 回应：⟨response⟩ 评分：</td>
</tr>
</tbody>
</table>

Table 3.4: Full text for **MEEP+SA** prompt translated into Spanish and Chinese; system role prompt in English, Spanish, and Chinese
<table>
<thead>
<tr>
<th>Prompt Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEEP + SA-DIAL</td>
<td>In this task, you will be shown part of a dialogue. Score the dialogue on a continuous scale from 0 to 100, where a score of zero means ‘disengaging’ and a score of 100 means ‘very engaging’. Consider that engagingness of each dialogue is defined by the following qualities: variety of responses according to the context (such as responding to ‘Hi how are you?’ with ‘I feel magnificent, because I just successfully defended my PhD! How are you? instead of ‘Good, how are you?’), likelihood of encouraging the other participant to respond (such as ‘I love legos! I like using them to make funny things. Do you like legos?’ instead of ‘I like legos.’), likelihood of encouraging quality responses from the other participant, interestingness, specificity, and likelihood of creating a sense of belonging for the other participant. Consider the overall engagingness of the conversation. Dialogue: ⟨ dialogue ⟩ Score: ⟨ score ⟩</td>
</tr>
</tbody>
</table>

Table 3.5: Full text for dialogue-level MEEP+SA-DIAL prompt. G-Eval prompts are edited slightly for applicability to dialogue-level evaluation.
4 Experiments and Results

We first quantitatively test all our prompts on an English-language dataset with turn-level annotations. A random selection of five examples from these tests are further evaluated qualitatively. Our best prompts are then used on six turn-level datasets – two each in English, Spanish, and Chinese – to comprehensively assess their performance in a multilingual setting. Our prompts are also translated into the language of the dataset being tested so performance can be assessed when the prompt language is the same as the language used in the dialogue. A final set of tests uses our best prompts on four dialogue-level datasets – two each in English and Chinese.

4.1 Benchmark Datasets

We use four existing datasets, all from the 11th Dialogue System Technology Challenge, Track 4 (DSTC11 [17]. Challenge organizers machine translated two English turn-level datasets into Spanish and Chinese, and two Chinese dialogue-level datasets into English to produce a total of ten datasets spanning three languages. Annotations from the original datasets are used to test the datasets in the additional languages. Details about each dataset are in Table 4.1.

- **FED** The Fine-grained Evaluation of Dialogue (FED) dataset [40] provides annotations of English-language dialogues first collected by [5] by way of multi-dimensional evaluations per dialogue turn, including engagingness as one of the evaluated dimensions. Each turn is given six scores for engagingness – one from each human evaluator who reviewed the dialogue. The instructions for human
<table>
<thead>
<tr>
<th>Dataset</th>
<th># samples</th>
<th>Language</th>
<th>Annotation level</th>
</tr>
</thead>
<tbody>
<tr>
<td>FED-EN</td>
<td>375</td>
<td>English</td>
<td>turn</td>
</tr>
<tr>
<td>FED-ES</td>
<td>375</td>
<td>Spanish</td>
<td>turn</td>
</tr>
<tr>
<td>FED-ZH</td>
<td>375</td>
<td>Chinese</td>
<td>turn</td>
</tr>
<tr>
<td>SEE-EN</td>
<td>300</td>
<td>English</td>
<td>turn</td>
</tr>
<tr>
<td>SEE-ES</td>
<td>300</td>
<td>Spanish</td>
<td>turn</td>
</tr>
<tr>
<td>SEE-ZH</td>
<td>300</td>
<td>Chinese</td>
<td>turn</td>
</tr>
<tr>
<td>KDCONV-EN</td>
<td>392</td>
<td>English</td>
<td>dialogue</td>
</tr>
<tr>
<td>KDCONV-ZH</td>
<td>392</td>
<td>Chinese</td>
<td>dialogue</td>
</tr>
<tr>
<td>LCCC-EN</td>
<td>22</td>
<td>English</td>
<td>dialogue</td>
</tr>
<tr>
<td>LCCC-ZH</td>
<td>22</td>
<td>Chinese</td>
<td>dialogue</td>
</tr>
</tbody>
</table>

Table 4.1: Dataset details including the number of samples used, language of the dialogues, and level at which the annotations were made.

evaluation of engagingness used in the creation of this dataset are used as the HD prompt.

- **SEE** The Persona-See (SEE) dataset [62] expands the PersonaChat task [88] to provide human annotations at the turn-level for multiple dimensions including enjoyment. Each dialogue is between a human and conversational model. At the conclusion, the human was asked how much they enjoyed the conversation. This is the score for enjoyment that we use. Due to the nature of the scoring, there is one score of enjoyment for each dialogue. The creators of the dataset consider enjoyment to be an accurate measure of engagement that users self-rate based on their active participation and positive assessment. We believe this to be an acceptable assumption. The dataset contains over 3,000 annotated turns. Since this is much larger than the other datasets used, and we would like to conserve resources and use comparable dataset sizes, we limit our testing to a random selection of 300 dialogues. We use the same 300 dialogues in each of the English-, Spanish-, and Chinese-language datasets.
• **KDCONV** The Knowledge-driven Conversation (KDCONV) dataset consists of Chinese-language dialogues between two humans who are given knowledge graphs from which to draw their responses [94]. Human evaluation of engagingness is annotated at the dialogue-level along with several other dimensions. These dialogues, originally in Chinese, were machine translated into English for Track 4 of the DSTC11 Challenge.

• **LCCC** The large-scale cleaned Chinese conversation dataset (LCCC) includes conversations from Chinese social media users who are believed to be human based on heuristic evaluation [76]. Their posts are parsed to create conversations, which are combined with dialogues from high-quality publicly available datasets. They are annotated by human evaluators for several qualities including engagingness. In the same process as for KDCONV, these dialogues were also translated into English for the DSTC11 Challenge.

4.2 LLMs

We use two families of large language models in our experiments.

• **GPT** For our base models, we select text-davinci-003 (GPT-3.5) [48], gpt-3.5-turbo-0301 (ChatGPT), and gpt-3.5-turbo-0613 (ChatGPT0613) from the GPT-series of LLMs [54]. We note that ChatGPT is available at the time of writing as gpt-3.5-turbo-* in the free version, and gpt-4 in the paid version. We use ChatGPT as the shorthand for gpt-3.5-turbo-0301, the free version at the time of testing. We set the temperature to 0, top_p to 1, and n to 1. Presence penalty and frequency penalty are set to 0, logit bias to null, and we request the best of 1.

• **LLaMA2** We also use LLaMA2 [69] in the 7B size with the chat_completion
function with \texttt{max\_seq\_len} set to 1024, \texttt{max\_gen\_len} set to None, \texttt{temperature} set to 0, \texttt{max\_batch\_size} set to 1, and \texttt{top\_p} set to 1. We do not set a max length for generation because the responses are long and the location of scores within the returned text is unpredictable.

4.3 Baselines

We compare the performance of our proposed metric against four state-of-the-art baselines.

- **EnDex** [82] uses fine-tuned RoBERTa-large models on datasets drawn from Reddit posts and manually evaluated at the turn-level. Evaluations measure dimensions of engagement including attentional, behavioral, emotional, and task-based qualities. Responses are rated as either engaging or not engaging. We were not able to adapt EnDex for use with languages other than English.

- **UniEval** evaluates any dimension, posing questions to a T5-large model [57]. Prompts are formed so they request a yes or no response. UniEval takes three inputs: the dialogue context, additional context, and the response. This is designed to work with popular public dialogue datasets such as Persona-See that include persona details. Since the FED dataset does not include an additional context, we pass in an empty string to take the place of the additional context required as a parameter. To the extent of our knowledge, this is the first evaluation of UniEval in a multilingual context.

- **GPTScore** uses the GPT family of models to evaluate the quality of dialogue response generation [24]. The metric evaluates eight dimensions of quality, including engagingness. A careful look at their code and description of methodology informed our use of their methodology. The prompt includes the task
definition, identification of the quality to be evaluated (in this case, engaging-ness), and the dialogue followed by ‘Answer: Yes’. The LLM is configured with temperature=0, max_tokens=0, logprobs=0, echo=True, and n=0. The log probabilities returned for the token ‘yes’ become the score for engagingness. The prompt for GPTScore is in Appendix C.

- **G-Eval** [36] is an LLM-based metric that uses Chain-of-Thought reasoning and a post-processing scoring function to evaluate dialogue response generation. It has strong correlations with human annotations using GPT-3.5 and GPT-4 models. Their highest correlations are with GPT-3.5 and we therefore use this version as the G-Eval baseline. We adapt their prompt to work with datasets that have only the dialogue context and response, omitting the portions referring to additional context. The prompt for G-Eval is in Appendix C.

### 4.4 Results of Translation Quality Analysis

Recall that some of the datasets have been translated into other languages by the original creators of the datasets. First, we investigate the quality of such translated datasets in this task. Translations into Spanish and Chinese were obtained from the MS Azure service and Tencent Machine Translation service by the organizers of the DSTC11 Challenge. The translations include four measures of translation quality for every utterance: COMET-20 quality estimator, COMET-21 quality estimator, CosSim1 cosine similarity, and CosSim2 cosine similarity. COMET-20 and COMET-21 are quality estimator metrics from the COMET models wmt20-comet-qe-da-v2 [59] and wmt21-comet-qe-mqm [58]. CosSim1 and CosSim2 measure cosine similarity of the original utterance and the translation after generating embeddings for both with the SentenceTransformer library [60]. They use the multilingual models
distiluse-base-multilingual-cased -v1, and paraphrase-xlm-r-multilingual -v1 respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>COMET-20</th>
<th>COMET-21</th>
<th>CosSim1</th>
<th>CosSim2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FED-EN</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>FED-ES</td>
<td>0.458</td>
<td>0.117</td>
<td>0.846</td>
<td>0.914</td>
</tr>
<tr>
<td>FED-ZH</td>
<td>0.323</td>
<td>0.084</td>
<td>0.789</td>
<td>0.883</td>
</tr>
<tr>
<td>SEE-EN</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SEE-ES</td>
<td>0.510</td>
<td>0.128</td>
<td>0.873</td>
<td>0.921</td>
</tr>
<tr>
<td>SEE-ZH</td>
<td>0.376</td>
<td>0.102</td>
<td>0.800</td>
<td>0.875</td>
</tr>
<tr>
<td>KDCONV-EN</td>
<td>0.242</td>
<td>0.053</td>
<td>0.788</td>
<td>0.857</td>
</tr>
<tr>
<td>KDCONV-ZH</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LCCC-EN</td>
<td>-0.132</td>
<td>-0.40</td>
<td>0.690</td>
<td>0.776</td>
</tr>
<tr>
<td>LCCC-ZH</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 4.2: Average machine translation quality scores of datasets that were translated. The highest scores for each measure are in **bold**. Second highest scores are underlined.

In order to understand the quality of the datasets as an accurate measure of engagingness, we aggregate the translation quality scores for each dataset in Table 4.2, showing that the Spanish-language translations score higher than the Chinese-language translations.

### 4.5 Results for Turn-level Evaluation

We evaluate the results from each phase of testing, considering the effect of each prompt style and performance in a multilingual context at the turn level and at the dialogue level. The effect of the language used for prompting is also evaluated.

#### 4.5.1 Results for Prompt Styles

Our naive prompt performs best on the simplest models – GPT-3.5 and LLaMA2-7B. The human-directed prompt yielded better performance than the naive prompt out of
Table 4.3: Results for prompts: Naive, Human-Directed (HD), and MEEP, with ablation results for prompt elements *such as* (SA) and system role (R). Spearman (S) and Pearson (P) results are listed for each experiment. Results marked with * are not statistically significant. The highest result for each column is in **bold**, with the second highest result underlined. The FED-EN dataset was used for testing.

<table>
<thead>
<tr>
<th>Prompt</th>
<th>GPT3.5 S</th>
<th>ChatGPT S</th>
<th>ChatGPT0613 S</th>
<th>LLaMA2-7B S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.445</td>
<td>0.521</td>
<td>0.434</td>
<td>0.416</td>
</tr>
<tr>
<td>Naive+R</td>
<td>0.416</td>
<td>0.466</td>
<td>0.464</td>
<td>0.471</td>
</tr>
<tr>
<td>HD</td>
<td>0.509</td>
<td>0.511</td>
<td>0.462</td>
<td>0.511</td>
</tr>
<tr>
<td>HD+R</td>
<td>0.492</td>
<td>0.500</td>
<td>0.497</td>
<td>0.518</td>
</tr>
<tr>
<td>MEEP</td>
<td>0.475</td>
<td>0.508</td>
<td>0.548</td>
<td>0.511</td>
</tr>
<tr>
<td>MEEP+R</td>
<td>0.494</td>
<td>0.507</td>
<td>0.540</td>
<td>0.514</td>
</tr>
<tr>
<td>MEEP+SA</td>
<td>0.532</td>
<td>0.520</td>
<td>0.558</td>
<td>0.514</td>
</tr>
<tr>
<td>MEEP+SA+R</td>
<td>0.526</td>
<td>0.517</td>
<td>0.568</td>
<td>0.514</td>
</tr>
</tbody>
</table>

the more advanced models and generally worse performance otherwise. HD prompts have average correlations of 0.420 (S) and 0.429 (P), the lowest correlations seen.

As seen in Table 4.3, the MEEP prompt shows improvement over the Naive prompt and the HD prompt, and results where the MEEP prompt were used are generally higher than results where it was not used. MEEP-based prompts have average correlations of 0.492 (S) and 0.481 (P), significantly better than those seen with the HD prompt.

In Table 4.3, we see that prompts including *such as* improve upon those without in 14 out of 16 instances. Prompts including *such as* have average correlations of 0.506 (S) and 0.492 (P). Their performance dominates both with and without defining the system role.

As seen in Table 4.3, the use of system role has mixed impact. Performance for ChatGPT and ChatGPT0613 improves with system role in most cases, while performance varies for LLaMA2-7B and generally falls for GPT-3.5. Defining the system role gives the most consistent gains in performance when used with ChatGPT
on prompts that do not include MEEP. GPT-3.5 almost entirely performs worse with the prepended system role.

4.5.2 Results for Models

In Table 4.3, we see that the models have different levels of performance, and that models can respond differently to each of the prompt styles we have tested, as noted above. LLaMA has the lowest performance with 0.151 (S) and 0.302 (P) lower average performance than GPT3.5. Both models have higher correlations with simpler prompts than the more advanced models.

We see that ChatGPT has an average of 0.023 (S) and 0.012 (P) higher correlations than GPT-3.5 across all experiments, and has similar though slightly higher aggregate performance to that of ChatGPT0613. Using our MEEP prompts, ChatGPT is the most aligned with human annotations.

4.5.3 Qualitative Analysis

Five randomly selected examples from prompt evaluation using the turn-level English FED dataset are shown in Tables 4.6 - 4.10. Although we see some confirmation in these examples with the aggregate data in Table 4.3, the patterns we see here do not consistently match up with those trends. With some exceptions, we see that MEEP prompts have higher correlation with human annotations, and ChatGPT performs better or similarly to GPT-3.5, two trends that we see in Table 4.3 as well.

Through the qualitative analysis, we notice that GPT-3.5 gives scores that are higher than the human annotations and hypothesize that this model consistently rates dialogue responses as being more engaging than other models. To check this, we average the scores for GPT-3.5 and ChatGPT and find that GPT-3.5 gives an
average score of 0.8138, while ChatGPT gives an average score of 0.7307.

### 4.5.4 Multilingual Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Prompt</th>
<th>FED EN</th>
<th>FED ES</th>
<th>FED ZH</th>
<th>SEE EN</th>
<th>SEE ES</th>
<th>SEE ZH</th>
</tr>
</thead>
<tbody>
<tr>
<td>EnDex</td>
<td>–</td>
<td>0.290</td>
<td>–</td>
<td>–</td>
<td>0.164</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UniEval</td>
<td>–</td>
<td>0.190</td>
<td>0.258</td>
<td>0.076*</td>
<td>0.015*</td>
<td>0.073*</td>
<td>0.038*</td>
</tr>
<tr>
<td>GPTScore</td>
<td>GPTScore</td>
<td>0.176</td>
<td>0.146</td>
<td>0.230</td>
<td>0.087*</td>
<td>0.153</td>
<td>0.140</td>
</tr>
<tr>
<td>G-Eval</td>
<td>G-Eval</td>
<td>0.488</td>
<td>0.448</td>
<td>0.402</td>
<td>0.194</td>
<td>0.131</td>
<td>0.062*</td>
</tr>
<tr>
<td>GPT-3.5</td>
<td>MEEP+SA</td>
<td>0.532</td>
<td>0.481</td>
<td>0.451</td>
<td>0.236</td>
<td>0.223</td>
<td>0.189</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>MEEP+SA</td>
<td>0.558</td>
<td>0.516</td>
<td>0.431</td>
<td>0.169</td>
<td>0.138</td>
<td>0.133</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>MEEP+SA+R</td>
<td>0.568</td>
<td>0.542</td>
<td>0.435</td>
<td>0.160</td>
<td>0.150</td>
<td>0.140</td>
</tr>
<tr>
<td>ChatGPT-0613</td>
<td>MEEP+SA+R</td>
<td>0.550</td>
<td>0.471</td>
<td>0.400</td>
<td>0.214</td>
<td>0.200</td>
<td>0.175</td>
</tr>
<tr>
<td>GPT-3.5</td>
<td>MEEP+SA†</td>
<td>–</td>
<td>0.438</td>
<td>0.520</td>
<td>–</td>
<td>0.128</td>
<td>0.085*</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>MEEP+SA†</td>
<td>–</td>
<td>0.500</td>
<td>0.408</td>
<td>–</td>
<td>0.161</td>
<td>0.149</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>MEEP+SA+R†</td>
<td>–</td>
<td>0.525</td>
<td>0.444</td>
<td>–</td>
<td>0.123</td>
<td>0.168</td>
</tr>
<tr>
<td>ChatGPT-0613</td>
<td>MEEP+SA+R†</td>
<td>–</td>
<td>0.542</td>
<td>0.374</td>
<td>–</td>
<td>0.273</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Table 4.4: Correlation results using Spearman coefficient for multilingual turn-level experiments. All results are statistically significant except those labeled with *. ‘†’ denotes a translated version of the prompt into the language of the dataset. Highest results are in **bold**.

In Table 4.4, we see the results for turn-level evaluation on six datasets – two each in English, Spanish, and Chinese – in which we compare our best prompts against four strong baselines. Our highest-performing English prompts have an average of 0.064 higher correlation than the highest-performing baseline metric for that dataset. In comparing the highest correlation seen for each language, improvement with our prompt over the next highest baseline is 0.060 for English datasets, 0.082 for Spanish datasets, and 0.050 for Chinese datasets.

For English-language prompting, correlations are highest for English-language
datasets, followed by Spanish- and Chinese-language datasets in that order. Results with prompts that were translated into the language of the dataset used for testing have the same or higher performance as the prompts in English. We see the highest results on the ChatGPT-0613 model. Additionally, results for the FED dataset are higher than results for the SEE dataset.

### 4.6 Results for Evaluation at the Dialogue-level

<table>
<thead>
<tr>
<th>Model</th>
<th>Prompt</th>
<th>KDCONV</th>
<th>LCCC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EN</td>
<td>ZH</td>
</tr>
<tr>
<td>G-Eval</td>
<td>G-Eval-DIAL</td>
<td>0.327*</td>
<td>0.189</td>
</tr>
<tr>
<td>GPT-3.5</td>
<td>MEEP+SA-DIAL</td>
<td>0.286*</td>
<td>0.223*</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>MEEP+SA-DIAL</td>
<td>0.178*</td>
<td>0.392*</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>MEEP+SA+R-DIAL</td>
<td>0.185*</td>
<td>0.428</td>
</tr>
<tr>
<td>ChatGPT-0613</td>
<td>MEEP+SA+R-DIAL</td>
<td>0.186*</td>
<td>0.362*</td>
</tr>
</tbody>
</table>

Table 4.5: Correlation results using Spearman coefficient for multilingual dialogue-level experiments. Results that are not statistically significant are marked with *. Highest scores for each dataset are in **bold**.

Table 4.5 reports results for dialogue-level evaluation in a multilingual setting on four datasets – two each for English and Chinese. We do not have many statistically significant results in this set of tests. We still see that our prompts outperform the baseline. Dialogue-level results are not as strong as turn-level results.
Dialogue: “Hi!”
“Hey there! How’s it going?”
“Hi, how are you?”
“I’m doing well, thank you. How are you?”
“I’m fine. What’s up?”
“Not much. How about you?”
“nothing new. Anything exciting happening recently?”
“Not really. You?”
“I watched a movie recently which was pretty good”
“That’s good. What movie?”
“the irishman, have you seen it?”
“I have not. Should I?”
“I think so”
“Alright, I’ll put it on my watch list.”
“How about you? Have you watched anything interesting recently?”
“I’ve been waiting to see the new Star Trek movie.”
“what is it about?”
“The adventures of Kirk, Spock, and McCoy.”
“when was it released?”
“It just came out today.”
“are you sure, I think it’s been out for a while”
“I’m positive. I know because I just watched it.”
“well, the fact you just watched it doesn’t mean it was released today”
“That’s what I said, I did watch it today.”
“ok, I’ll put it on my watchlist”
“Let me know what you think of it when you watch it.”
“sure, I need to go, nice talking to you”
Response: “Good talking to you too.”

Normalized average of human annotations: 0.4

<table>
<thead>
<tr>
<th>Prompt</th>
<th>GPT-3.5</th>
<th>δ</th>
<th>ChatGPT</th>
<th>δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.8</td>
<td>0.4</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>Naive + R</td>
<td>0.8</td>
<td>0.4</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>HD</td>
<td>0.0</td>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>HD+R</td>
<td>0.5</td>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>MEEP</td>
<td>0.7</td>
<td>0.2</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>MEEP+R</td>
<td>0.7</td>
<td>0.2</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>MEEP+SA</td>
<td>0.7</td>
<td>0.2</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>MEEP+SA+R</td>
<td>0.7</td>
<td>0.2</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 4.6: Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth (δ) for each score is listed to the right of the score.
Dialogue: “Hi!”
“Hi, how’s it going?”
“All good! It’s cold”
Response: “It’s pretty cold here too.”

Normalized average of human annotations: 0.7

<table>
<thead>
<tr>
<th>Prompt</th>
<th>GPT-3.5</th>
<th>δ</th>
<th>ChatGPT</th>
<th>δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.8</td>
<td>0.1</td>
<td>0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>Naive + R</td>
<td>0.8</td>
<td>0.1</td>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td>HD</td>
<td>0.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>HD+R</td>
<td>0.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>MEEP</td>
<td>0.8</td>
<td>0.1</td>
<td>0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>MEEP+R</td>
<td>0.8</td>
<td>0.1</td>
<td>0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>MEEP+SA</td>
<td>0.7</td>
<td>0.0</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>MEEP+SA+R</td>
<td>0.7</td>
<td>0.0</td>
<td>0.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 4.7: Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth (δ) for each score is listed to the right of the score.
Dialogue: “Hi!”
“Hey!”
“How’s your day going?”
“Having fun, but have a bit of a headache How about you?”
“I’m good! Just getting ready to watch a basketball game. Hope your headache goes away!”
“Haha, thanks! Never get them, so it’s super weird Who’s playing?”
“Hmm, try taking some REDACTED_TERM! I’m going to see the Warriors vs. the Grizzlies. Ever see them play?”
“I haven’t Honestly I probably wouldn’t be able to identify which sport they played”
“Haha, all the animal names confuse me as well! Who came up with these naming conventions? I’d name my team something more original: The New York Chairs!, sounds better to me :) I’m actually at the REDACTED_TERM opening remarks”
“Oh cool! What is ‘NeurIPS’?”
“Neural Information Processing Systems”
“Oh wow! Sounds fancy! What is that about?”
“It’s the biggest ML conference in the world”
“Oh, ML?”
Response: “Yes, machine learning”

<table>
<thead>
<tr>
<th>Prompt</th>
<th>GPT-3.5</th>
<th>δ</th>
<th>ChatGPT</th>
<th>δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.8</td>
<td>0.0</td>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Naive + R</td>
<td>0.8</td>
<td>0.0</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>HD</td>
<td>0.0</td>
<td>0.8</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>HD+R</td>
<td>0.0</td>
<td>0.8</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>MEEP</td>
<td>0.4</td>
<td>0.4</td>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>MEEP+R</td>
<td>0.7</td>
<td>0.1</td>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td>MEEP+SA</td>
<td>0.4</td>
<td>0.4</td>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td>MEEP+SA+R</td>
<td>0.4</td>
<td>0.4</td>
<td>0.7</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 4.8: Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth (δ) for each score is listed to the right of the score.
Dialogue: “Hi!”
“Have you seen any good movies lately?”
“Last weekend I saw ‘The Parasite’. Ever heard of it?”
Response: “No. Why did you pick that movie?”

Normalized average of human annotations: **0.9**

<table>
<thead>
<tr>
<th>Prompt</th>
<th>GPT-3.5</th>
<th>δ</th>
<th>ChatGPT</th>
<th>δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.8</td>
<td>0.1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>Naive + R</td>
<td>0.8</td>
<td>0.1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>HD</td>
<td>1.0</td>
<td>0.1</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>HD+R</td>
<td>1.0</td>
<td>0.1</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>MEEP</td>
<td>0.8</td>
<td>0.1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>MEEP+R</td>
<td>0.8</td>
<td>0.1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>MEEP+SA</td>
<td>0.8</td>
<td>0.1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>MEEP+SA+R</td>
<td>0.8</td>
<td>0.1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 4.9: Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth ($\delta$) for each score is listed to the right of the score.
**Dialogue:** “Hi!”

“Hi! How are you?”

“I am good”

... 

“That’s good! What are you up to today?”

“I am trying to get my work done”

“How’s that going so far?”

“like is fun”

“That’s good! What are you supposed to be doing right now?”

“Helping collect some data for training models.”

“Ooh, what kind of training models?”

“Models that will help improve tools for our users.”

“Ooh, that sounds like a good job!”

“Thank you, that’s for sure”

“You’re welcome! How long until you’re done with training models?”

“We keep on trying to improve them so I guess it’ll be a while”

**Response:** “That’s understandable. Good luck, I hope it goes smoothly!”

**Table 4.10:** Comparison of results for one example dialogue from the FED dataset for qualitative analysis. The distance from the ground truth ($\delta$) for each score is listed to the right of the score.
5 Discussion

This chapter discusses the results of our evaluation and their implications in detail.

5.1 Interpretation of Results

The better performance of the naive prompt on GPT-3.5 and LLaMA2-7B may be because the more detailed prompts are difficult for the smaller models to follow. From a qualitative perspective, the LLaMA2-7B results were often delivered in a form that could not easily be parsed and had to be manually interpreted. It appeared that the model did not process the requested format for its response and it is possible that this limitation also affected the quality of the scores given. This may be why the naive prompt produces better results on the simpler models.

The complexity of the human-directed prompt may have contributed to its worse performance on the simpler models. When used with the larger models, the specificity that the complexity allows may be what yields better performance on the larger models than the naive prompt. Despite this improvement, HD prompts have the lowest average correlations of any prompts. It is important to remember that the HD prompt used here is the same that was used in the creation of the dataset used for testing. It therefore is significant that this is not as successful as other prompting methods. We can infer that the way natural language is processed in these models is different from the way it is processed by humans.

The higher correlations seen with our MEEP prompt indicates that using our definition of engagingness improves the quality of evaluation produced by these models.

The inconsistent improvements with the use of system role may be because it
offers a contextual shift that is not as closely aligned as other methods, like MEEP. This more approximate alignment at times moves the model to a less-aligned position than it would be if using a better aligned method alone. The system role prompt is not beneficial when used with GPT-3.5 in this setting. This, combined with mixed performance when used on LLaMA2-7B suggests that it may add a level of complexity to which the simpler models cannot rise.

The higher average scores given by GPT-3.5 than those give by ChatGPT confirms our hypothesis that GPT-3.5 has a positive bias in comparison with ChatGPT and we consider that this may be an indicator of the model’s limitations.

The lower correlations of our MEEP prompts on Spanish- and Chinese-language datasets is not what we would expect given that both Spanish and Chinese are high-resource languages. We theorize that it is related to the translation quality of the Spanish and Chinese datasets which in Table 4.2, we confirm is lower for Chinese than for Spanish.

Because we see better performance with prompts translated into the language of the dialogues, we infer that this combination gives better results. Higher results on the ChatGPT-0613 model suggests that improved system role capabilities with this model may have included multilingual system role training. We consider that translation of dialogues may affect the validity of the engagingness scores with which they are paired, while potentially lowering the quality from their versions in the original languages. We suggest that there may be a cultural component to this, and that engagingness may be seen differently for speakers of English, Chinese, or Spanish. In fact, there can be many subcultures within these languages and engagingness can be seen very differently in each of these subcultures.

We note that the results for the FED dataset may be higher because each response has six scores for engagingness, whereas each response in the SEE dataset has only
Although the scores given in the FED dataset are from a third-person perspective, the effect of having multiple people scoring the responses may give a more reliable aggregate score, and mitigate the noise. Low correlations for dialogue-level results may indicate that these models are better equipped to evaluate engagingness at the turn level.

5.2 Contributions

Key findings of this research are as follows:

- Our MEEP dimensions of engagingness improve alignment with human annotations and are an effective component of an LLM-based metric, improving over state-of-the-art metrics in evaluation of engagingness. Clear improvement is seen with the use of ‘such as’ when used with clarifying examples in this context and we conclude that examples written in this form will improve dialogue evaluation performance with these LLMs.

- The results for our MEEP prompts used on multilingual datasets show improvement over state-of-the-art baselines in the evaluation of engagingness in dialogue across Chinese, Spanish, and English. This is also consistent across turn-level and dialogue-level datasets.

- The language of the prompts affects which model will give the most human-aligned evaluation. The model that performs best with MEEP in English is ChatGPT. When used with MEEP translated into Spanish or Chinese to match the language of the dialogues, the model that performs the best is ChatGPT-0613.

- In the multilingual setting, for turn-level dialogues, automatic evaluation is
most strongly correlated with human evaluation when the prompt is in the language of the dialogues.

- Model size affects the type of prompt that yields the best results. ChatGPT performs better than GPT-3.5 with more complex prompts. When prompts are in their simplest form, as with the Naive or HD prompts, GPT-3.5 has higher correlation with human annotations. LLaMA gives its highest correlations when used with the naive prompt. We infer that to see an increase with the more powerful model, the context must be appropriately set, as with MEEP+SA, and that simpler prompts are more appropriate for smaller LLMs.

- The effect of role assignment is uneven when compared as used with the set of prompts studied here. Defining the system role offers some improvement when used with ChatGPT or ChatGPT0613 with prompts that are naive. It does not improve performance when used with LLaMA2-7B, or GPT-3.5. While role assignment doesn’t generally improve performance, it also was linked to our best overall scores. The non-linear nature of the effect of role assignment points to the complexity of designing a metric with LLMs.
6 Conclusion

We successfully demonstrate the potential of Large Language Models (LLMs) for automatically evaluating the engagingness of dialogue in multilingual settings. Through the development of a nuanced definition of engagingness and the implementation of tailored prompting techniques, we have surpassed state-of-the-art automatic metrics and also show promise for further exploration into the evolution of dialogue system evaluation.

Our research establishes a foundational understanding of engagingness as a multifaceted construct that encompasses elements such as response diversity, interactional quality, interestingness, contextual specificity, and a sense of belonging. By leveraging the capabilities of LLMs, we have shown that it is feasible to evaluate these aspects in dialogues across multiple languages, addressing the critical need for scalable and efficient evaluation tools in the rapidly evolving domain of conversational AI.

Looking ahead, several directions for future research emerge from this study:

1. **Culture-specific system prompt**: Use of the system prompt to instruct the LLM to identify as a member of a specific cultural group may affect its ability to evaluate engagingness of dialogues between people within that cultural group. Exploring this possibility could show the effect of cultural perspectives on setting the context for defining engagingness.

2. **Effect of length of dialogue**: Testing differences in length of dialogues and correlations with human annotations could show the effect of dialogue length on LLMs' ability to evaluate.
3. **Vary examples given with ‘such as’**: We acknowledge the possibility that the examples given with the ‘such as’ statements may introduce a bias to the evaluation of the responses and dialogues. Exploration of results with different cultural contexts and comparable or contrastive examples included could elucidate the biases implied with these examples.

4. **Evaluation of consistency**: Although our results are largely statistically significant, the consistency of our results has not been tested. An evaluation of these results with multiple runs of each test would give an evaluation of how consistent results are and how reliable the metric is.

5. **Further exploration of dimensions of engagingness**: Our five dimensions of engagingness were subjectively defined based on existing information drawn from multiple sources. A more robust method of identification of dimensions of engagingness may further improve performance of our metric.

6. **Evaluation of the correlation of LLMs with each other**: It is still unknown how the LLMs’ performance correlates with other LLMs. Further analysis could provide insights into their performance by doing correlation analysis on the results seen for each of the LLMs.

7. **Re-scoring engagingness after translation of dialogues**: The use of scores from dialogues in their original languages for translated dialogues is unexplored as a source of noise in our results. Re-scoring the dialogues for engagingness after translation may provide more accurate scores that can be better predicted by an automatic metric. Performing this re-scoring may give results that are able to provide deeper insights about the role of translation in multilingual evaluation of dialogues.
8. **In-depth Analysis of Engagingness Across Diverse Contexts**: Engagingness, being inherently subjective, varies significantly across different social, cultural, and conversational contexts. Future research should aim to dissect these variances by conducting comprehensive studies that explore how engagingness is perceived and manifested in diverse contexts. By analyzing dialogues from varied cultural backgrounds and conversational settings, we can develop more robust and universally applicable evaluation metrics.

9. **Enhancement of LLMs for Low-Resource Languages**: The current success in evaluating engagingness across multiple languages should be extended to include low-resource languages, which are often underrepresented in conversational AI research. Future efforts should be directed toward developing datasets and enhancing LLM capabilities in these languages. This will not only broaden the applicability of engagingness evaluation but also contribute to the inclusivity and diversity of conversational AI technologies.

10. **Ethical Considerations and Bias Mitigation**: As LLMs take on a larger role in evaluating dialogue engagingness, it is crucial to address the ethical implications and potential biases inherent in these models. Future research should explore the ethical dimensions of automated engagingness evaluation, identifying biases and developing methodologies to mitigate them. This will take steps toward making evaluation processes less biased and more reflective of the diverse range of human interactions.

11. **Evaluation of Other Complex Conversational Qualities**: This research shows the promise of LLMs for evaluation of dialogue agents for complex qualities like inclusivity, emotional tone, or humor. We see potential for future research in exploring evaluation of these qualities.
12. **Use of Efficient Models**: Future research could explore methods to achieve accurate engagingness evaluation with smaller, more efficient models. Use of models like Mixtral 8x7b, known for its good performance and efficiency in multilingual settings, would be a useful area of future research. This could reduce resource consumption and make the evaluation process more accessible and sustainable.

We have taken a significant step toward understanding and evaluating the engagingness of dialogue using LLMs. The proposed future research directions offer a path toward further advancement, with promise for developing more nuanced, inclusive, and ethically responsible evaluation tools for dialogue systems. As conversational AI continues to evolve, we hope that this work can contribute to the improvement of dialogue systems ability to engage users more effectively and meaningfully.
Bibliography


[38] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. Chatgpt as a factual inconsistency evaluator for text summarization, 2023.


[61] John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam


## A Parameters and Performance of Dialogue Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>Open Source</th>
<th>MMLU %</th>
<th>MT Bench</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-1</td>
<td>1.5 million</td>
<td>Yes</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GPT-2</td>
<td>1.5 billion</td>
<td>Yes</td>
<td>32.4</td>
<td>–</td>
</tr>
<tr>
<td>GPT-3</td>
<td>175 billion</td>
<td>No</td>
<td>53.9</td>
<td>–</td>
</tr>
<tr>
<td>GPT-3.5</td>
<td>not disclosed</td>
<td>No</td>
<td>70.0 (few-shot)</td>
<td>–</td>
</tr>
<tr>
<td>GPT-3.5 Turbo</td>
<td>not disclosed</td>
<td>No</td>
<td>–</td>
<td>7.94</td>
</tr>
<tr>
<td>GPT-4</td>
<td>not disclosed</td>
<td>No</td>
<td>86.5</td>
<td>8.99</td>
</tr>
<tr>
<td>GPT-4 Turbo</td>
<td>not disclosed</td>
<td>No</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gemini Nano-1</td>
<td>1.8 billion</td>
<td>No</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gemini Nano-2</td>
<td>3.25 billion</td>
<td>No</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gemini Pro</td>
<td>not disclosed</td>
<td>No</td>
<td>79.13 CoT@8</td>
<td>–</td>
</tr>
<tr>
<td>Gemini Ultra</td>
<td>not disclosed</td>
<td>No</td>
<td>90.04</td>
<td>–</td>
</tr>
<tr>
<td>LLaMA1 7b</td>
<td>7 billion</td>
<td>Yes</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LLaMA1 13b</td>
<td>13 billion</td>
<td>Yes</td>
<td>–</td>
<td>2.61</td>
</tr>
<tr>
<td>LLaMA1 33b</td>
<td>33 billion</td>
<td>Yes</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LLaMA1 65b</td>
<td>65 billion</td>
<td>Yes</td>
<td>63.4</td>
<td>–</td>
</tr>
<tr>
<td>LLaMA2 7b</td>
<td>7 billion</td>
<td>Yes</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LLaMA2 13b</td>
<td>13 billion</td>
<td>Yes</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LLaMA2 70b</td>
<td>70 billion</td>
<td>Yes</td>
<td>68.9</td>
<td>–</td>
</tr>
</tbody>
</table>

Table A.1: Popular LLMs, number of parameters, whether they are open source, and scores on MMLU and MT Bench – two commonly used metrics.
B LLM Architecture

B.1 Transformers

At the core of LLMs is the transformer architecture [71]. This architecture at a high-level is an encoder-decoder. The concept of the encoder-decoder is that a set of mathematical processes creates an encoded version of the input. The encoding can then be decoded to generate the desired result. In a Convolutional Neural Network (CNN) architecture, the encoding can be a smaller version of the input, that can be conceptualized as a distillation of the essential information in the input. The encoding is then expanded again by a decoder which, in the CNN example, creates a full-sized output based on the distillation created by the encoder. The compression and expansion steps are the encoding and decoding that happens in an encoder-decoder. This allows for something new to be generated based on the input. In image processing, it can be a sharpened image. In natural language processing, it can be a translation, summarization, or response to a question.

The transformer architecture has many encoder and decoder units, as seen in Figure B.1. The encoders are strung together so the output of one is fed into the next. The decoders are also strung together, taking as input the output from the previous decoder. The final output from the string of encoders is also passed to each of the decoders.
Figure B.1: The transformer architecture. Diagram by Yuening Jia [31].
B.2 Tokenization and positional embedding

To begin processing word-based input into a transformer architecture, the input words are tokenized, and an embedding is produced from a look-up table. The embedding is then added to a positional encoding to be processed in the first encoder.

B.3 Encoding

Each encoder unit contains a self-attention layer followed by a feed-forward layer. Self-attention gives a relevance rating for the relationship between each of the tokens in the input. The feed-forward layer processes each word through an identical neural network to produce another set of embeddings. At each of these steps, the input is added to the output and normalized before inputting into the next step. After the final encoder, the embeddings are transformed into two attention vectors to be passed to the decoders.

B.4 Decoding

Decoder structure is similar to that of the encoder, with an added attention layer between self-attention and feed-forward layers. This encoder-decoder attention layer takes the two attention vectors from the encoding process as two of its three inputs to inform where the decoder focuses as it processes its third input, the output from the self-attention layer. This self-attention layer is slightly different from the one in the encoders in that it is masked to allow scoring of only relationships between the current word and previous words.

The decoder section is run multiple times until an end of output token is generated by the decoder. Each time this section is run, a word is produced and this becomes
the input for the next run of the decoders. As with the beginning of the encoding section, each token of this input is transformed into an embedding through a look-up table and summed with its positional encoding before it is processed by the first decoder unit.

After the final decoder unit, the output is processed through a linear layer to produce logits, which are soft-maxed to get probabilities for each word in the vocabulary used. The highest-probability token is selected. In another option, beam search saves the top n tokens, runs the decoding with each of these tokens, and the final output is selected from these sets of options.

B.5 Training

During training, a labeled dataset is used and the difference between the predicted word and the actual word in the dataset is measured through a loss function. Through back-propagation, this loss is used to adjust the weights on the attention, feed-forward, and linear layers in each of the encoder and decoder units.

B.6 Fine-Tuning

A model trained on a dataset based on one context will not perform as well when the context changes. This can be overcome through fine-tuning in which a relatively small dataset specific to a new context is used to provide continued training (fine-tuning) on a model that has already been trained on a more general or less-related dataset. This is a powerful way to expand the utility of LLMs.
B.7 One-shot or Few-shot Training

A minimalized version of fine-tuning, one-shot or few-shot training, inputs one or a few examples into a language model in order to improve performance [11,77].

B.8 Prompting

Structuring input into LLMs through prompt engineering can give improved performance [77,96]. Beyond one-shot and few-shot training, examples of prompt engineering techniques include structuring a prompt to request multi-step reasoning, follow detailed instructions, predicting intermediate outputs as with a scratchpad, and self-evaluation [77].

B.9 Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) is used to further fine-tune a transformer [48,97]. This involves several steps: collecting a dataset produced by hired response generators; fine-tuning the transformer on the new dataset; creating a comparative dataset with multiple responses to a single prompt ranked by human evaluators; using this comparative dataset to train a reward model; and using the reward model to further train the transformer. RLHF is used extensively in training LLMs.

B.10 Mixture of Experts

In order to provide deeper expertise in multiple contexts, an architecture called Mixture of Experts (MoE) [8] trains multiple networks on specialized datasets. A gating function selects the weights with which the outputs from each of the specialized net-
works will contribute to the final output [21].

B.11 Achieving State of The Art Performance

Many variations exist on how to integrate transformer architecture, fine-tuning, RLHF, and MoE for optimal performance [64]. Most top-performing models’ architectures and training methodologies are not shared with the public, but are thought to combine these methods, along with increasing the number of parameters and size of training data sets in order to optimize performance.
## C Baseline Metric Prompts

<table>
<thead>
<tr>
<th>Prompt Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPTScore</strong></td>
<td>Answer the question based on the conversation between a human and AI. Question: Are the responses of AI engaging? (a) Yes. (b) No. Dialogue context: ⟨ conversation ⟩ Response: ⟨ response ⟩ Answer: Yes.</td>
</tr>
<tr>
<td><strong>G-Eval</strong></td>
<td>You will be given a conversation between two individuals. You will then be given one potential response for the next turn in the conversation. Your task is to rate the responses on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Engagingness (1-3) Is the response dull/interesting? - A score of 1 (dull) means that the response is generic and dull. - A score of 2 (somewhat interesting) means the response is somewhat interesting and could engage you in the conversation (e.g., an opinion, thought) - A score of 3 (interesting) means the response is very interesting or presents an interesting fact Evaluation Steps: 1. Read the conversation, the corresponding fact and the response carefully. 2. Rate the response on a scale of 1-3 for engagingness, according to the criteria above. 3. Provide a brief explanation for your rating, referring to specific aspects of the response and the conversation. Example: Conversation History: ⟨ dialogue ⟩ Response: ⟨ response ⟩ Evaluation Form (scores ONLY): - Engagingness:</td>
</tr>
</tbody>
</table>

Table C.1: Full text for GPTScore and G-Eval prompts.
<table>
<thead>
<tr>
<th><strong>Prompt Code</strong></th>
<th><strong>Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>G-EVAL-DIAL</td>
<td>You will be given a conversation between two individuals. Your task is to rate the conversation on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Engagingness (1-3) Is the conversation dull/interesting? - A score of 1 (dull) means that the conversation is generic and dull. - A score of 2 (somewhat interesting) means the conversation is somewhat interesting and could engage you in the conversation (e.g., an opinion, thought) - A score of 3 (interesting) means the conversation is very interesting or presents an interesting fact Evaluation Steps: 1. Read the conversation carefully. 2. Rate the conversation on a scale of 1-3 for engagingness, according to the criteria above. 3. Provide a brief explanation for your rating, referring to specific aspects of the conversation. Example: Conversation: ⟨ dialogue ⟩ Evaluation Form (scores ONLY): - Engagingness:</td>
</tr>
</tbody>
</table>

Table C.2: Full text for dialogue-level G-EVAL-DIAL prompt. G-Eval prompts are edited slightly for applicability to dialogue-level evaluation.