ConvKGYarn: Spinning Configurable and Scalable Conversational Knowledge Graph QA datasets with Large Language Models

Ronak Pradeep[1,2], **Daniel Lee**[1,3], **Ali Mousavi**[1], **Jeff Pound**[1], **Yisi Sang**[1], **Jimmy Lin**[2], **Ihab Ilyas**[1], **Saloni Potdar**[1], **Mostafa Arefiyan**[1] and **Yunyao Li**[4*]

[1] Apple [2] University of Waterloo [3] University of Calgary [4] Adobe

rpradeep@uwaterloo.ca {mostafaa, s_potdar}@apple.com
yunyaol@adobe.com

Abstract

The rapid advancement of Large Language Models (LLMs) and conversational assistants necessitates dynamic, scalable, and configurable conversational datasets for training and evaluation. These datasets must accommodate diverse user interaction modes, including text and voice, each presenting unique modeling challenges. Knowledge Graphs (KGs), with their structured and evolving nature, offer an ideal foundation for current and precise knowledge. Although human-curated KG-based conversational datasets exist, they struggle to keep pace with the rapidly changing user information needs. We present ConvKGYarn, a scalable method for generating up-to-date and configurable conversational KGQA datasets. Qualitative psychometric analyses confirm our method can generate high-quality datasets rivaling a popular conversational KGQA dataset while offering it at scale and covering a wide range of human-interaction configurations. We showcase its utility by testing LLMs on diverse conversations — exploring model behavior on conversational KGQA sets with different configurations grounded in the same KG fact set. Our results highlight the ability of ConvKG-Yarn to improve KGQA foundations and evaluate parametric knowledge of LLMs, thus offering a robust solution to the constantly evolving landscape of conversational assistants.

1 Introduction

The proliferation of Large Language Models (LLMs) and conversational assistants has led to their ubiquitous presence in daily user interactions. This widespread adoption underscores the critical need for dynamic datasets capable of rigorously evaluating their proficiency in addressing knowledge-seeking queries. Knowledge Graphs (KGs) have long been recognized as powerful tools for capturing structured representations of the world (Hogan et al., 2021). In KGs, concepts and entities are represented as nodes, while semantic relationships defining facts are represented with edges. This structured representation has had an impact across various domains, including Natural Language Processing (Schneider et al., 2022), Recommender Systems (Guo et al., 2022), and Information Retrieval (Reinanda et al., 2020).

The integration of LLMs and KGs has opened up new opportunities in natural language processing (Petroni et al., 2019; Guu et al., 2020; Peng et al., 2023), which has led to significant advancements across various tasks (Barba et al., 2021; Chakrabarti et al., 2022; De Cao et al., 2022; Xu et al., 2023). By combining the dynamic capabilities of LLMs with the structured insights from KGs, researchers have unlocked new avenues for developing advanced question-answering (QA) systems. In conversational Knowledge Graph Question Answering (KGQA), datasets like ConvQuestions (Christmann et al., 2019) have emerged to address scenarios where questions often lack full context or contain grammatical inconsistencies. These datasets have played a crucial role in enabling new retrieval-augmented systems, demonstrating the potential of LLM-KG integrations to provide accurate and attributable responses in conversational settings (Christmann et al., 2023).

In relation, advancements in text retrieval have underscored the potential of using LLMs to generate synthetic data to improve the effectiveness of downstream systems. This process has been utilized at scale for neural query synthesis (Nogueira and Lin, 2019; Ma et al., 2022; Pradeep et al., 2022) and LLM-based ranked list reorderings for instruction distillation into open-source rerankers (Pradeep et al., 2023a,b; Tamber et al., 2023) resulting in substantial improvements across a spectrum of retrieval tasks. More recently, synthetic data generation, facilitated by automated prompt optimization (Xian et al., 2024), has en-

[*]Work done while at Apple

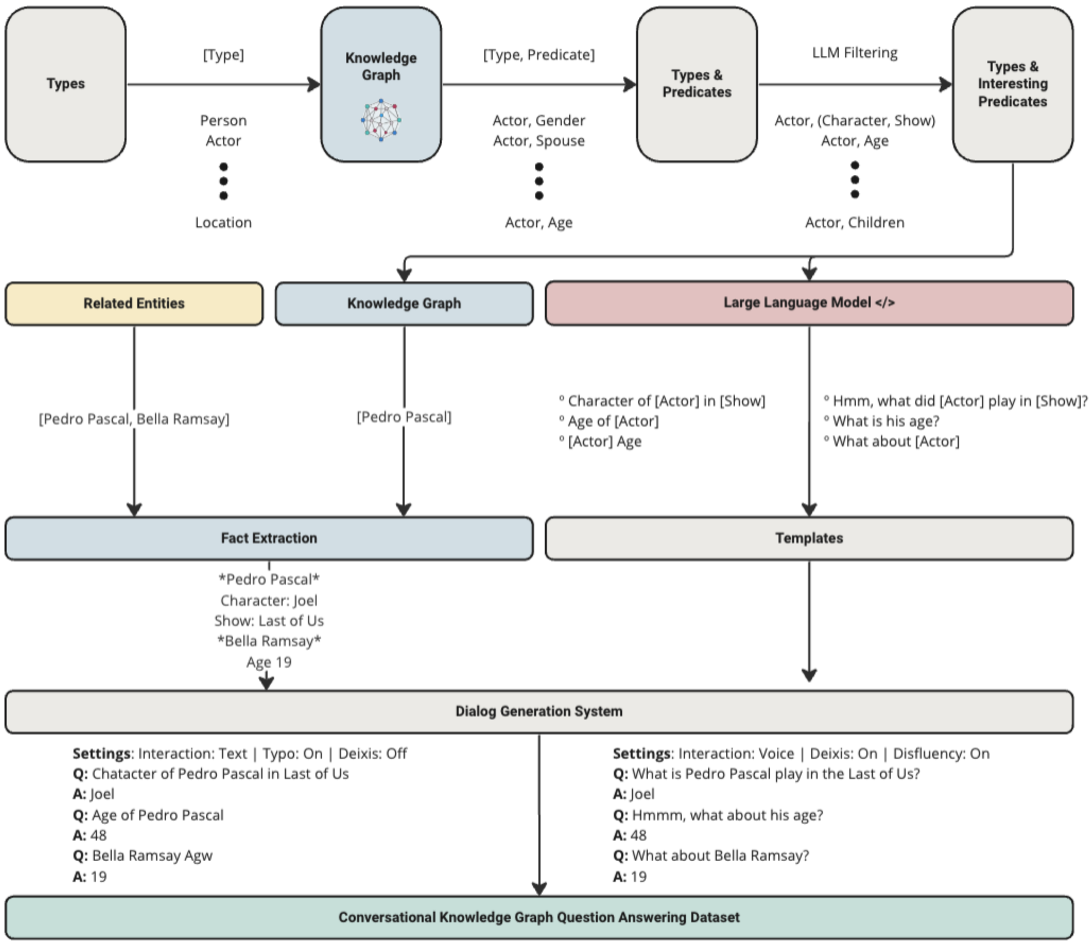


Figure 1: The full ConvKGYarn pipeline.

abled the training of highly effective small-scale models, deprecating the dependence on human-labeled data. Collectively, these highlight the opportunity to integrate synthetic data strategies from LLMs to develop more resilient and adaptable knowledge-intensive systems.

While existing conversational KGQA datasets are rich in content, they often struggle to keep pace with rapidly evolving user information needs. This discrepancy raises questions about the relevance of such data in real-world, adaptive conversational scenarios. To address this challenge, we introduce ConvKGYarn, a novel method for generating large-scale, configurable conversational Knowledge Graph Question Answering (KGQA) datasets. Through rigorous psychometric evaluation metrics, we demonstrate that ConvKGYarn produces high-quality conversational data comparable to established, human-curated KGQA datasets. Notably, ConvKGYarn achieves this while significantly expanding entity and fact coverage by several orders of magnitude and introducing configurable properties in user interaction styles.

A critical component of our research involves evaluating the datasets generated by ConvKGYarn using various LLMs to assess their parametric knowledge. Our observations reveal that these models often struggle with fact recall, underscoring the need for retrieval-augmented systems. By evaluating datasets generated with different interaction styles and their varied linguistic phenomena generated with consistent fact sets from the Knowledge Graph, we aim to assess the robustness of LLMs in handling diverse conversational settings, in confounder-free manners. Our findings indicate that model effectiveness varies significantly across diverse user interaction styles, highlighting the importance of developing LLMs capable of serving as robust conversational systems.

Through this work, we seek to reveal the path toward creating datasets that can effectively train and evaluate evolving conversational assistants. We envision these methods and datasets will play a crucial role in developing more versatile and adaptive conversational AI systems.

2 ConvKGYarn

Figure 1 illustrates the entire pipeline of the ConvKGYarn system. We first introduce the key notations and definitions that are of use in our framework. Following this, we dive into each module that comprises the ConvKGYarn system.

2.1 Definitions and Notations

The knowledge graph (KG) acts as the foundation for our work. Following the Wikidata terminology, we have an item (or entity), $e \in \mathcal{E}$ and statements (or facts) $\mathcal{S}_e$, that we represent by *item-property-value* tuples that help describe the fact. The properties (or predicates) are denoted by $p_\mathrm{e} \in \mathcal{P}_\mathrm{e}$.

We denote values (or objects) for a particular entity and predicate p_e, with o_{p_e}. In ConvKGYarn, we use a *simple fact* to denote item-property-value tuples where the corresponding predicate does not involve multiple entries. Conversely, some entities naturally possess properties with multiple values, such as the siblings of an Actor or the official languages of a Country. These are acceptable within ConvKGYarn as *complex facts*. Another exception, which we dub *qualified facts*, is where different values hold in light of qualifiers (for example, the population of a Country or the CEO of a company can both include qualifiers by timestamps). These qualifiers further delineate or refine the values within a statement and are accepted in the ConvKGYarn framework.

Each entity e is associated with multiple *types* T_e with a specific type denoted by t_e. Examples of types are Person, Singer, Book, and Movie. Note that in ConvKGYarn, in addition to using the `InstanceOf` predicate to describe types, we also leverage the `Occupation` predicate to add nuances to the types, and from there, the predicates we draw our attention to since we believe that the *interesting* predicates for someone who is a Politician are likely different from one who is an Actor.

2.2 KG Predicate Extraction

The initial stage of ConvKGYarn leverages the KG to extract all predicates p_i for a particular entity type T. This extraction process is denoted by $\mathcal{F}(t) = \{p_1, p_2, \ldots, p_n\}$, where $\mathcal{F}$ is the extraction function, t is a type, and $\{p_1, p_2, \ldots, p_n\}$ is the set of predicates such that there exists some entity e of type t, for which p_i is a valid predicate. In our example in Figure 1, we see that for an Actor, these include predicates like Gender or Spouse.

2.3 LLM Predicate Selector

This step employs a large language model (LLM) to filter the extracted predicates, selecting only the most *interesting* predicates for each entity type. This process is governed by the prompt detailed in Figure 2 in Appendix A and can be formulated as $\mathcal{G}(t, \{p_1, p_2, \ldots, p_n\}) = \{p'_1, p'_2, \ldots, p'_m\}$, where $\mathcal{G}$ represents the selector function, selecting a subset $\{p'_1, p'_2, \ldots, p'_m\}$ from the initial set of predicates.

By prompting with a high degree of specificity, we hope the selection optimizes for predicates that enhance the richness of the dataset while maintaining contextual appropriateness. Additionally, by including the Wikidata identifier of the predicate, we hope to resolve cases where the identifier name is unclear, especially given that these models have most likely encountered them during training.

Furthermore, we expect the predicates that pass through this filter to contribute meaningfully to conversations surrounding the entity type. To achieve this, we prompt the model to exclude predicates that are too generic, irrelevant noise, or identifiers, none of which lend themselves to a high-quality conversational QA dataset in any interaction form.

We see in Figure 1 that for an Actor, ConvKGYarn selects predicates like (Character, Show), a qualified predicate, or Age.

2.4 Related Entity Generator

The related entity generator $\mathcal{R}$ is an additional component of ConvKGYarn that identifies and selects entities e_r linked to the primary entity e. Doing this allows for the enrichment of the dataset with diverse but relevant information that is often not directly in the vicinity of the original entity (for example, as seen in Figure 1, actors like Pedro Pascal and Bella Ramsay might not be direct neighbors on Wikidata graph, yet questions about them could show up in the same conversation by their association through the Last of Us TV series). Related entities can be selected using KG embedding similarity (inner product) with embeddings that prioritize capturing the ontology of the graph. We use only the *most-similar* related entity for popular Person entities to not introduce bias or excessive noise into our datasets.

2.5 Fact Extraction

Using the KG, ConvKGYarn extracts factual information $\mathcal{I}$ corresponding to each entity. For an entity e, we represent the fact extraction for simple or complex facts by $\mathcal{I}(e) = \{(e, p'_1, o_1), \ldots, (e, p'_m, o_m)\}$, where o_i denotes the object(s) corresponding to the "interesting" predicate p'_i.

In the case of *qualified facts*, we can generalize this to include $\mathcal{I}_c(e) = \bigcup_{i=1}^{m} \bigcup_{j=1}^{l_i} \{(e, p'_i, q_i, o_i)\}$, where q_i is the qualifier set.

2.6 Synthetic Question Template Generation

To maintain configurability and scalability, in addition to ensuring the tractability of ConvKGYarn framework, we generate questions using a templated approach that incorporates placeholder entity type (with actual type information in natural language form, for example, [actor]), interesting predicates, and placeholder objects ([i]) in the prompt. For simple and complex facts, the detailed prompt for generating questions for voice interactions and textual (or search) interactions are in Figure 3 and Figure 4 in Appendix A, respectively. The prompt for qualified facts is in Figure 5 (in Appendix A).

In designing ConvKGYarn, it was imperative to emulate the nuances of both text and voice interactions, representing the primary modalities through which users engage with AI assistants. The goal was to capture the essence of these interactions with the prompts, spotlighting the differences in user experience. In text interactions, we aim to mimic search queries, emphasizing short keyword queries and with follow-ups made in succession. These interactions allow for *deixis*, where questions reflect references to previously mentioned entities in the conversation, enhancing their continuity. Additionally, ConvKGYarn also has a knob for typographical errors (typos), another phenomenon common in textual interactions, albeit in a post-processing step discussed in Section 3. In voice interactions, we hope the system generates more well-formed questions. In addition to *deixis*, the voice modality allows for conversations with *disfluencies*. Disfluencies aim to mimic the imperfections of natural speech by adding natural uses of *uh*, *um*, takebacks, apologies, thanks, or repetitions, among others. We hope to amalgamate these aspects in the "deixis_disfluencies" variants to simulate the intricacies of human conversation, involving both references and speech errors.

The structured prompt ensures that for each fact and linguistic phenomenon, we generate three question variants. Doing so ensures more variations in the generated questions versus sampling questions by querying the LLM multiple times, which is slower, more expensive, and less guaranteed to output variants. Additionally, by generating all variants together, we hope to have variants with linguistic phenomena that build on the same original variant. This approach helps better evaluate the robustness of conversational QA systems or LLMs while providing comprehensive training data that includes a wide range of linguistic variations.

We speed up inference by providing five triples instead to save on the bulk of the input tokens (the instructions). Note that the turn number does not mean ConvKGYarn is generating a question for that specific turn, although they do capture sequential order important for *qualified facts*. Instead, its purpose is to give an index for both the JSON key and the object identifier.

The JSON format used in the prompt is pivotal for systematic data parsing during the generation process. It ensures that the questions are generated in a consistent format, facilitating easy integration into the rest of our pipeline.

To allow qualified facts in ConvKGYarn, we generalized the standard triples to tuples with the additional relational predicate field. Note that while turn-specific objects are disallowed in the questions, objects from other turns that belong to the same predicate are encouraged to help create more complex questions. For example, the query "voice of [a] in [movie]" could correspond to turn 2 of the prompt with the answer "[b]".

These configurations within the prompt are designed to maximize the effectiveness and applicability of the synthetic questions, making them fundamental to generating realistic and varied conversational QA instances. Accounting for linguistic variability and contextual appropriateness enables ConvKGYarn to curate robust, scalable, and highly configurable conversational KGQA datasets.

2.7 Conv. Factoid QA Instance Creation

Finally, a subset of extracted facts for an entity e, along with those for its related entities (if they exist), can be slot-filled using examples from the generated templates to get a conversation instance. Note that this instance creation step adheres to some rules. Regardless of the interaction type and

Dataset	# Entities	# Facts	# Unique Types	# Unique Predicates	# Questions Per Fact
General	29M	196M	274	1252	24
Related	210K	6.1M	95	265	54

Table 1: Dataset statistics for the various settings considered.

selected linguistic phenomena, the first turn never involves any deixis. We group certain predicates to ensure cohesiveness and avoid weird artifacts in the final conversations. For instance, questions about date of birth or place of birth are likely to occur near each other instead of being separated by several facts.

This process combines fresh, up-to-date factual data from the KG with synthetic templates, examinable by humans, to form a factoid KGQA instance. Given that templatized generation and slot-filling are significantly cheaper than generating specific conversations for each new entity, ConvKGYarn allows us to curate large-scale, configurable datasets efficiently.

3 Experimental Setup

In all our experiments, we utilize a Wikidata dump with a knowledge cut-off date of June 2023. To ensure consistency within our pipeline, which assumes English-language input, we filter the dump to only include entities with English names. The original Wikidata dump contained approximately 100 million entities. However, after a meticulous cleaning pipeline and filtering to include only English examples and entities of *interesting* types, our final dataset comprises 29 million entities associated with 196 million facts.

For the LLM predicate selector, we employed the GPT_4 model, utilizing the `gpt-4-0613` endpoint provided by OpenAI. To maintain the model's reasoning efficacy and prevent cognitive overload, we limit our LLM requests to a maximum of 50 predicates at a time. We implement this selection process in a segmented manner, ensuring comprehensive coverage of each type–predicate pair. For predicates with linked qualifiers, we include the relationship predicate in the input to provide the necessary context. This methodical approach allows us to isolate predicates that are particularly relevant for conversational factoid QA.

For the generation of synthetic question templates, we utilize the `gpt-3.5-turbo` endpoint. We provide two in-context examples for each prompt (omitted for length) to better align generations to the expected template format.

For textual interactions, while calling the model endpoint, we leverage the "logit_bias" field to penalize the model when it generates one of the question words — wh-words or how. Without doing this, we found the model ignores instructions and in-context examples to generate fully-formed questions instead.

For typo augmentation, going over each turn's question, we select at random one of the following attacks part of the `TextAttack` framework (Morris et al., 2020): `WordSwapRandomCharacterDeletion()`, `WordSwapNeighboringCharacterSwap()`, or `WordSwapQWERTY()` and apply it. We introduce a single "meaningful" typo to each question turn.

4 Dataset Statistics

Table 1 presents high-level statistics of the two large sets of data we curated using ConvKGYarn.

The *General* set encompasses all entities and their associated facts from our filtered Wikidata, without incorporating any notion of related entities. This comprehensive collection comprises 29 million entities and an extensive 196 million facts. For each fact, our methodology generates 24 possible questions: 12 from voice interactions (three each from original, deixis, disfluencies, and deixis_disfluencies sets), and 12 from textual interactions (three each from original, deixis, typos, and deixis_typos sets). This approach exponentially increases the potential for generating diverse conversations, providing a large-scale resource for training conversational agents and exposing large language models to high-quality synthetic data. The dataset's complexity and realism are further enhanced by its inclusion of 274 unique types and 1,252 unique predicates. This level of scale and coverage is challenging to achieve and is not typically observed in human-curated datasets. For instance, ConvQuestions (Christmann et al., 2019) contains only 11,200 real-user conversations, with an average of five questions each, derived from just five primary entity types.

The *Related* set, while more specialized than

	Model	Interaction	Deixis	Disfluency	Typo	Fluency	Relevance	Diversity	Grammar	Agreement
(1)	ConvKGYarn$_G$	Voice	✗	✗	-	3.97	4.63	2.40	3.90	75.5
(2)	ConvKGYarn$_G$	Voice	✗	✓	-	3.39	4.49	2.25	3.37	73.5
(3)	ConvKGYarn$_G$	Voice	✓	✗	-	3.99	4.59	2.45	3.77	74.8
(4)	ConvKGYarn$_G$	Voice	✓	✓	-	3.29	4.41	2.32	3.02	71.0
(5)	ConvKGYarn$_R$	Voice	✗	✗	-	3.70	3.71	2.66	3.69	68.5
(6)	ConvKGYarn$_R$	Voice	✗	✓	-	3.34	3.74	2.59	3.39	67.6
(7)	ConvKGYarn$_R$	Voice	✓	✗	-	3.76	3.89	2.79	3.72	69.3
(8)	ConvKGYarn$_R$	Voice	✓	✓	-	3.36	3.73	2.73	3.38	71.5
(9)	ConvKGYarn$_G$	Text	✗	-	✗	2.83	4.41	2.19	2.95	70.8
(10)	ConvKGYarn$_G$	Text	✗	-	✓	2.61	4.36	2.17	2.18	68.8
(11)	ConvKGYarn$_G$	Text	✓	-	✗	2.84	4.36	2.29	2.83	67.1
(12)	ConvKGYarn$_G$	Text	✓	-	✓	2.29	4.09	2.00	1.63	73.0
(13)	ConvKGYarn$_R$	Text	✗	-	✗	2.57	3.38	2.58	2.75	66.3
(14)	ConvKGYarn$_R$	Text	✗	-	✓	2.29	3.45	2.45	1.97	71.5
(15)	ConvKGYarn$_R$	Text	✓	-	✗	2.48	3.33	2.54	2.73	70.8
(16)	ConvKGYarn$_R$	Text	✓	-	✓	2.12	3.31	2.58	1.86	68.5

Table 2: The results from the Single Model Rating of the *General* (ConvKGYarn$_G$) and *Related* (ConvKGYarn$_R$) set reflecting Likert scores of 1-5 for Fluency, Relevance, Diversity, and Grammar. Agreement scores represent the mean percentage of all scores where at least two of three annotators agree.

the *General* set, offers a focused exploration of popular Human-type entities, comprising 210,000 entities associated with 6.1 million facts. Despite its smaller scale, it provides a higher density of question variants, with an average of 54 questions per fact. This includes the 24 questions generated for the *General* set, plus an additional 30 questions derived from related entity-specific follow-up queries, enabling a more detailed and nuanced exploration of topics. Encompassing 95 unique types and 265 unique predicates, this targeted dataset facilitates in-depth exploration and evaluation of conversational systems focused on human-centric entities. The specialized structure of the *Related* set makes it particularly valuable for assessing the effectiveness of AI systems in handling complex, interconnected queries about popular entities.

5 Results

To evaluate the efficacy of ConvKGYarn, we employ three methods to comprehensively understand its quality and usefulness: (1) Single-Model Rating, (2) Pairwise Comparison, and (3) Parametric Knowledge Evaluation of LLMs.

Combining these methods aims to complement the other's strengths and weaknesses. Likert scores, typically used in single-model grading, while very scalable as a human evaluation method, have several inherent limitations when used to evaluate language models. It relies on annotators making absolute judgments rather than relative comparisons, which tends to be less reliable for humans (Stewart et al., 2005). The result can be inconsistent biases between different annotators (Kulikov et al., 2019). While pairwise comparisons avoid some of these issues by having annotators make relative judgments between pairs of data points, comparing a set of models is less efficient, often requiring re-evaluation of existing baseline models whenever a new model is introduced (Stewart et al., 2005). Finally, we explore how LLMs fare at the synthetically generated conversational factoid QA datasets generated by ConvKGYarn, investigating their fact recall abilities by leveraging LLM-as-a-Judge evaluation and adding a critical dimension to our story. While this scales better while often correlating strongly with human annotations, they still suffer from issues like the self-enhancement bias, where LLM may favor the answers generated by themselves (Zheng et al., 2023). Together, these methods provide a robust and multifaceted approach to thoroughly evaluating the efficacy of ConvKGYarn, ensuring a comprehensive assessment from both human and automated perspectives.

5.1 Single-Model Rating

The Single-Model Rating task presents human annotators with a multi-turn conversation, in which they assign a score from 1-5 across four parameters.

We evaluated the dataset across four key parameters: Fluency, Relevance, Diversity, and Grammar. We assessed these parameters for 16 different combinations of settings available in the ConvKGYarn pipeline across 1600 conversations sampled with a uniform distribution, including Interaction (Voice or Text), Deixis (On or Off), Disfluency (On or Off, only for Voice), Typo (On or Off, only for Text), and Related Entities (On or Off). At a high level, we designed it to cover a diverse set of unique entities sampled from Wikidata, featuring a wide range of entity types such as Person, Actor, Singer, and Politician.

The task interface, which we designed on an internal annotation tool, and in-depth guidelines are in Appendix B, including a detailed explanation of the crowdsourcing process (onboarding, training, and quality).

Table 2 analyzes the alignment of the scores for each parameter with the 16 setting combinations. The introduction of typographical errors affects fluency and grammar quality as their presence can disrupt the smooth flow and grammatical accuracy of the conversations. Conversely, the inclusion of deixis can result in better fluency by creating more natural and contextually grounded conversations. However, deixis also impacts grammar, as referential expressions may introduce ambiguity or inconsistency in the dialogue structure.

As we would prefer, relevance and diversity appear resistant to variations in deixis, disfluencies, typographical errors, and interaction settings. The finding suggests that the content and informational diversity of the conversations remain largely unaffected by these factors. However, related entities impact the scores for relevance and diversity. By incorporating information from related entities, the conversations exhibit improved relevance to the topic and offer a wide range of knowledge discovery through the traversal of connected concepts in the KG.

Our analysis reveals that the optimal combination of settings for ConvKGYarn involves the voice interaction type with deixis and related entities. This configuration generates conversations that resemble natural human speech and discourse patterns, as reflected in the corresponding evaluation scores, minus disfluencies. Despite the inherent subjectivity of human evaluation, the conversations generated by ConvKGYarn exhibit an average annotator agreement of 70.53%, indicating a good level of consensus in their assessments and supporting the reliability of our evaluation.

These findings highlight the importance of considering multiple dimensions and interaction setting combinations when evaluating the quality of synthetically generated conversational datasets. By systematically exploring the impact of different factors on key evaluation parameters, we can gain a more nuanced understanding of the strengths and limitations of the ConvKGYarn approach. This knowledge can inform future pipeline refinements to enhance the quality and naturalness of the generated conversations.

Type	Fluency (%)	Relevance (%)	Diversity (%)	Grammar (%)
Preference	45.0	62.2	56.0	56.6
Agreement	84.6	89.0	82.2	86.6

Table 3: The results from the Pairwise Comparison. We indicate pairwise comparisons through Preference, i.e., the percentage of graders who prefer ConvKGYarn.

Since the datasets curated by ConvKGYarn feature a diverse set of synthetically curated conversational QA instances and cover various entity types, linguistic phenomena, and interaction modalities, our benchmark can comprehensively evaluate a model's ability to handle the nuances and challenges of real-world conversations.

It is critical to note that single-model dialogue grading may be affected by a lack of relative understanding compared to other datasets and curation methods.

5.2 Pairwise Comparison

The Pairwise Comparison task introduces human annotators to two conversations: (1) a conversation generated by ConvKGYarn and (2) a commonly used conversational KGQA dataset. For these two conversations, the annotators indicate their preferences across the same psychometric evaluation metrics outlined in Section 5.1 across 500 conversations focusing on voice interaction, without disfluencies and related entities.

The reference conversational QA dataset, ConvQuestions (Christmann et al., 2019), was chosen based on its similarity to ConvKGYarn's purpose and capabilities while being human-curated.

The dataset generated with ConvKGYarn adapted the process outlined in Section 5.1 with three changes to mirror the attributes of the benchmark dataset: we restrict entity types to the ones included in the benchmark dataset, use the entity referred to in the first turn of the reference conversation as the starter entity in ConvKGYarn, and ensure the number of turns of both datasets is equal.

ConvKGYarn demonstrates varying qualitative preference across the psychometric schema compared to human-curated reference conversations as shown in Table 3. In terms of fluency, ConvKGYarn nearly reaches parity with a preference of

	Model	Interaction	Deixis	Disfluency	Typo	Mean (*Turn*)	Mean (*Conv.*)	NA Ratio
(1)	$GPT_{3.5}$	Voice	✗	✗	-	0.246 / 0.326	0.234 / 0.323	0.485 / 0.304
(2)	$GPT_{3.5}$	Voice	✗	✓	-	0.250 / 0.349	0.236 / 0.346	0.434 / 0.272
(3)	$GPT_{3.5}$	Voice	✓	✗	-	0.261 / 0.305	0.244 / 0.303	0.440 / 0.312
(4)	$GPT_{3.5}$	Voice	✓	✓	-	0.261 / 0.306	0.254 / 0.304	0.432 / 0.276
(5)	$GPT_{3.5}$	Text	✗	-	✗	0.246 / 0.333	0.233 / 0.329	0.459 / 0.276
(6)	$GPT_{3.5}$	Text	✗	-	✓	0.220 / 0.279	0.199 / 0.277	0.513 / 0.352
(7)	$GPT_{3.5}$	Text	✓	-	✗	0.239 / 0.307	0.221 / 0.302	0.445 / 0.306
(8)	$GPT_{3.5}$	Text	✓	-	✓	0.201 / 0.220	0.179 / 0.219	0.519 / 0.433
(9)	GPT_4	Voice	✗	✗	-	0.301 / 0.391	0.292 / 0.387	0.352 / 0.252
(10)	GPT_4	Voice	✗	✓	-	0.320 / 0.412	0.307 / 0.407	0.329 / 0.232
(11)	GPT_4	Voice	✓	✗	-	0.299 / 0.374	0.288 / 0.370	0.333 / 0.269
(12)	GPT_4	Voice	✓	✓	-	0.299 / 0.384	0.290 / 0.381	0.340 / 0.244
(13)	GPT_4	Text	✗	-	✗	0.316 / 0.371	0.294 / 0.366	0.335 / 0.285
(14)	GPT_4	Text	✗	-	✓	0.265 / 0.347	0.242 / 0.346	0.451 / 0.350
(15)	GPT_4	Text	✓	-	✗	0.269 / 0.361	0.248 / 0.355	0.385 / 0.309
(16)	GPT_4	Text	✓	-	✓	0.222 / 0.290	0.201 / 0.285	0.479 / 0.396

Table 4: The effectiveness based on the GPT_4-EVAL metric of two models $GPT_{3.5}$ and GPT_4 when evaluated against variants of the *General* and *Related* settings (scores separated by /). Note that all these settings are grounded on the same set of facts.

45. 0%. This slight underperformance can be attributed to the writing style generated by language models, which at times may not realistically represent natural writing patterns. Additionally, inter-turn questions could lack the same cohesiveness as human-written conversations, given that we prioritize scale.

However, ConvKGYarn achieves significantly higher relevance with a 62.2% preference. We believe this is due to the methodology employed in ConvKGYarn, which generates questions from a diverse knowledge base encompassing the primary and related entities. In contrast, human-curated conversations rely on annotators researching the given entity to create the dialogues, potentially leading to higher variability and divergence from the initial entity.

ConvKGYarn shows modest improvements in diversity and grammar, with preference rates of 56.0% and 56.6% respectively. The slight advantage in grammar may be due to the standardized dialect and writing style of the LLM utilized in ConvKGYarn, compared to the inherent variance across human annotators. The diversity improvement, while notable, still leaves room for enhancement. This limitation likely stems from the structured method of generating questions based on entity types and relationships in the KG, which could constrain the range of topics compared to the more open-ended human curation process.

It's worth noting that the agreement among human annotators is exceedingly high for all ratings, ranging from 82.2% to 89.0%. This strong consensus lends credibility to the evaluation results and suggests a clear differentiation between ConvKGYarn-generated and human-curated conversations across the assessed dimensions.

Overall, the human evaluation of ConvKGYarn reveals that it surpasses or closely reaches parity with human-curated conversations across four key dimensions: fluency, relevance, diversity, and grammar. These findings challenge the common perception that synthetically generated datasets are inherently of lower quality. Instead, ConvKGYarn presents itself as a promising approach for generating high-quality conversational data in a repeatable and scalable manner.

5.3 Quantitative Analysis — Parametric Knowledge Evaluation

We explore the effectiveness of LLMs on 100 examples from each of the *General* and *Related* sets. For each set, ConvKGYarn generates conversations spanning *all* configurations considered. This grounding on a consistent fact set enables us to test different hypotheses in a confounder-free manner, allowing us to carefully analyze how LLMs fare specifically with typos or a combination of deixis and disfluences in the voice interaction setup.

Figure 6 in Appendix A presents an example interaction of how we evaluate LLMs on the conversational sets, iteratively as we go through each interaction turn of the conversational dataset. For each turn, we prepend the model with the *gold* *conversational history*. The prompt is designed to ensure that the model provides the most accurate and relevant information directly about the query, omitting extraneous details. In addition, the model can return responses in list form when multiple valid responses exist, ensuring clarity in the presentation of information. Finally, the prompt allows the return of "NA" if low in confidence, i.e., if it believes the knowledge captured in its parameters or the ambiguity in the question results in it being unable to answer the question accurately. The datasets were tested with two LLMs, $GPT_{3.5}$ and GPT_4.

Upon curating factoid answers from these models, we employ GPT_4 as a judge to rate the predictions in a binary fashion, as depicted in Figure 7 in Appendix A. Our evaluation prompt systematically assesses the correctness of responses. Each candidate answer is compared against the gold answer for each conversational turn. A score of 1 is assigned if the candidate addresses the query as per the gold standard; otherwise, a 0. Finally, we instruct the model to score list answers with a score of 1 if some candidate string matches any gold answer. The entire process respects the order of the conversation, providing scores in a list format that directly corresponds to the sequence of turns.

The results from this evaluation setup provide quantifiable metrics on the effectiveness of the tested LLMs, especially given that a metric like F1 and EM fails to accurately account for various aliases and other variations in LLM answers.

Table 4 presents the GPT_4-EVAL results for the variants from the *General* and *Related* settings. The metrics include the mean score assigned at a turn or conversation level. Additionally, we also report the rate of refusals (NA Ratio).

Firstly, we note that in both tables, GPT_4 scores higher across the board than $GPT_{3.5}$, rows (9)–(16) vs. (1)–(8). This finding could be due to the enhanced capabilities of the larger models, which also benefit from more extensive training data and refined instruction fine-tuning. The drop in the refusals is also indicative of GPT_4 having successfully stuffed a lot more information in the model parameters than the smaller $GPT_{3.5}$.

Second, we can see that it is inconclusive whether conversations in the voice interaction setting achieve higher scores than those in the textual interaction setting when not compounded by other linguistic phenomena, as illustrated in rows (1) vs. (5) and (9) vs. (13). In the presence of deixis, the outcomes are similarly nuanced, although slightly favoring the voice interaction setting, rows (3) vs. (7) and (11) vs. (15). This result suggests that these models can more easily resolve referents in spoken queries, which often contain more contextual clues than text-based keyword queries.

Third, the introduction of disfluencies, unexpectedly, appears to have a negligible or even slightly beneficial effect on the results in voice interaction settings, rows (2) vs. (1), (4) vs. (3), (10) vs. (9), and (12) vs. (11). These findings indicate that LLMs are becoming increasingly adept at filtering out irrelevant signals to focus on the core informational need of a query.

Finally, typos, as one might predict, diminish the effectiveness of both models. This is reflected in the decrease in scores whenever typos are present, rows (6) vs. (5), (8) vs. (4), (14) vs. (13), and (16) vs. (15), underscoring the models' sensitivity to correct spelling as a factor in understanding and processing questions.

Overall, these results provide a nuanced understanding of LLMs in the domain of conversational factoid question answering across diverse configurable settings. We posit that a comprehensive evaluation encompassing this array of configurations is imperative to develop a thorough portrayal of system effectiveness.

6 Conclusions

In this paper, we introduce ConvKGYarn, a novel framework for generating dynamic and scalable conversational datasets for Knowledge Graph Question Answering (KGQA). Our system leverages the structured representation of Knowledge Graphs (KGs) to produce configurable and adaptive conversational datasets that evolve with user information needs and KG-captured knowledge. Extensive evaluations demonstrate ConvKGYarn's effectiveness in generating very high-quality KGQA datasets. Through rigorous qualitative and quantitative tests, we showcase the versatility of these datasets across various conversational scenarios, enabling the assessment of models' effectiveness in different facets of user interactions and linguistic phenomena.

ConvKGYarn enhances the testing capabilities of LLMs and QA systems in adapting to the evergrowing knowledge landscape but also facilitates high-quality evaluation across different forms of user interactions, each with its nuances.

Acknowledgement

We would like to thank Anil Pacaci, Simone Conia and Varun Embar for insightful discussions surrounding the choices made during the project.

References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Soumen Chakrabarti, Harkanwar Singh, Shubham Lohiya, Prachi Jain, and Mausam . 2022. Joint completion and alignment of multilingual knowledge graphs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2023. Compmix: A benchmark for heterogeneous question answering. *arXiv:2306.12235*.

Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19. Association for Computing Machinery.

Simone Conia, Min Li, Daniel Lee, Umar Minhas, Ihab Ilyas, and Yunyao Li. 2023. Increasing coverage and precision of textual information in multilingual knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*.

Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2022. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Comput. Surv.*

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics.

Xueguang Ma, Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2022. Document expansions and learned sparse lexical representations for ms marco V1 and V2. In *Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022)*.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docttttquery.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv:2302.12813*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Ronak Pradeep, Yilin Li, Yuetong Wang, and Jimmy Lin. 2022. Neural query synthesis and domain-specific ranking templates for multi-stage clinical trial matching. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22. Association for Computing Machinery.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. RankVicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv:2309.15088*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv:2312.02724*.

Ridho Reinanda, Edgar Meij, and Maarte de Rijke. 2020. Knowledge graphs: An information retrieval perspective. *Foundations and Trends® in Information Retrieval*.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Neil Stewart, Gordon D A Brown, and Nick Chater. 2005. Absolute identification by relative judgment. *Psychol. Rev.*

Manveer Singh Tamber, Ronak Pradeep, and Jimmy Lin. 2023. Scaling down, litting up: Efficient zero-shot listwise reranking with seq2seq encoder-decoder models. *arXiv:2312.16098.*

Jasper Xian, Saron Samuel, Faraz Khoubsirat, Ronak Pradeep, Md Arafat Sultan, Radu Florian, Salim Roukos, Avirup Sil, Christopher Potts, and Omar Khattab. 2024. Prompts as Auto-Optimized Training Hyperparameters: Training best-in-class IR models from scratch with 10 gold labels. *arXiv:2406.11706*.

Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, and Dilek Hakkani-Tür. 2023. KILM: Knowledge injection into encoder-decoder language models. *arXiv:2302.09170*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv:2306.05685*.

A Additional Prompts

In this section, we include a few prompts that we could not include in Section 2 because of space restrictions. Figure 2 illustrates the prompt for predicate filtering. For simple and complex facts, the detailed prompt for generating templatized questions for voice and textual (or search) interactions are in Figure 3 and Figure 4, respectively. For qualified facts, we provide the prompt used in Figure 5.

Figure 6 presents an example interaction of how we evaluate LLMs on the conversational sets, iteratively as we go through each interaction turn of the conversational dataset. Upon curating factoid answers from these models, we employ GPT_4 as a judge to rate the predictions in a binary fashion, as depicted in Figure 7.

B Human Annotation Process

In this section, we provide in-depth details on ConvKGYarn's human annotation process used during the evaluation tasks. The human annotation interface and associated guidelines are attached in the appendix.

B.1 Psychometric Evaluation

The objective of the annotation process was to grade the provided conversation on a Likert scale of 5, across a defined psychometric evaluation schema. First, given a conversation, the human annotators were asked to familiarize themselves with its information: the user interface for the task provided a short overview of the instructions, as well as the evaluation schema upon which the conversation would be graded. In addition, the annotators were provided with a thorough instruction file, which correlated directly to the annotation task and gave granular details on the task, the evaluation schema, and helpful tips.

After learning about the task, the annotators were tasked with grading the conversation across the provided evaluation schema on a scale of 1 to 5. To do so, human annotators were recommended to become thoroughly familiar with the context of the conversation. The evaluation schema consisted of several psychometric dimensions, each with its own set of criteria and definitions. For each dimension, annotators could choose one of the following general options. However, the definition and scaling explanation was tailored to each dimension, to provide a granular understanding.

- 1 - Poor. The conversation fails to meet the criteria for the given dimension and exhibits significant issues or deficiencies.
- 2 - Fair. The conversation partially meets the criteria for the given dimension but has some notable weaknesses or areas for improvement.
- 3 - Satisfactory. The conversation adequately meets the criteria for the given dimension, with no major strengths or weaknesses.
- 4 - Good. The conversation effectively meets the criteria for the given dimension and demonstrates some notable strengths or positive qualities.
- 5 - Excellent. The conversation fully meets or exceeds the criteria for the given dimension, exhibiting exceptional quality or performance.

Annotators were given the choice to opt out from rating a conversation if they felt they did not have enough context or knowledge about the topic to make an informed assessment.

Please refer to the Dialogue Grading - Task Guidelines for further information on the evaluation schema and their definitions.

B.2 Comparative Analysis

Similar to the previous annotation task, the objective of this annotation process was to compare two conversations with a similar context, under the same psychometric evaluation schema. The task undertaken by the human annotators was the main difference between the two annotation processes.

First, given a pair of conversations, the human annotators were asked to familiarize themselves with the information provided: the user interface for the task presented a short overview of the instructions, as well as the evaluation schema upon which the conversations would be compared. In addition, the annotators were provided with a thorough instruction file, which correlated directly to the annotation task and gave granular details on the task, the evaluation schema, and helpful tips.

After learning about the task, the annotators were tasked with comparing the two conversations across the provided evaluation schema. The evaluation schema consisted of several psychometric dimensions, each with its own set of criteria and definitions. For each dimension, annotators could choose one of the following options:

SYSTEM: You are a helpful assistant that can help select all predicates likely to be used in a Factoid Conversational QA dataset for a particular type of entity. You should not select something like id/index/phone number/Commons category (which does not lend well to Conversational QA), name (which is obvious from the question itself), and also things which have little or nothing to do with the particular type like goals scored for a type actor or supported sports team for a singer. Predicates whose corresponding objects have type video, audio, and image should also not be included. Do not include first name and last name which would already be obvious from the user question. Things like marriage/partners should be included. You will be provided with a type and a table of tuples of the form (predicate_id, predicate_name). Always provide only an answer and in the format <pythonic list of useful predicate ids>.
USER: Type: *singer*
Predicates: [('P412', 'voice type'), ('P4431', 'Google Doodle'), ('P793', 'significant event'), ...]
GPT$_4$: *[('P412', 'voice type'), ...]*

Figure 2: Prompt for the LLM-based Predicate Selector.

- Conversation A. The first conversation better meets the criteria for the given dimension compared to the second conversation.
- Conversation B. The second conversation better meets the criteria for the given dimension compared to the first conversation.
- Same. Both conversations equally meet the criteria for the given dimension, with no significant differences between them.

Please refer to the Dialogue Comparisons - Task Guidelines for further information on the evaluation schema and their definitions.

B.3 Quality Assurance and Inter-Annotator Agreement

Closely adapted from Conia et al. (2023), to ensure the highest quality output, all human annotators were required to pass a rigorous entrance test before participating in the annotation process. This test involved studying a comprehensive set of guidelines that familiarized the annotator with the fundamental concepts of conversational KGQA, outlined the task and UI elements, and provided illustrative examples. Additionally, annotators had to successfully complete qualification exams tailored to each specific task, achieving a pre-defined threshold compared to the gold labels. Only annotators who passed the entrance test were permitted to proceed with the actual annotation process (the 25 conversations used in the entrance test were excluded from the final dataset).

We exclusively recruited annotators who could demonstrate proficiency in English, and limited the locales to either en-US or en-CA. Compensation for annotators was based on the competitive hourly wages per annotator's geographic location. On average, annotators dedicated approximately 5 minutes to each conversation. Given that each conversation was evaluated by 3 annotators, we estimate the total human time invested in the annotation process to be 3 annotators $\times$ 1,000 conversations $\times$ 5 minutes / 60 minutes = 250 hours.

Upon completion of the annotation process, we assessed inter-annotator agreement using a majority vote calculation. Table 4 illustrates an average agreement of 70.53% (Psychometric Evaluation) and 85.6% (Comparative Analysis) which is generally considered to be a strong level of agreement.

This inter-annotator agreement score serves to validate the results obtained from the annotation process.

SYSTEM: You are an AI assistant tasked with generating a natural conversational question-answering session between two people, A and B, based on information from a knowledge graph, in the form of a list of triples. A will only ask questions, and they should be based on the subject type and predicate of each triple, while B will only answer with just the object and no extraneous information. To make the conversation more realistic, you should also include for A:
- deixis (words that refer to people, places, or things in the conversation history like this, their, that, it, they, them)
- disfluencies (pauses, repetitions, and other speech errors that occur naturally in conversation)
- deixis_disfluencies (each question displays both deixis and disfluencies)
You only return JSON of the following form with key being an <int representing the turn number> mapping to:
- original: <list of three variants of standard single-turn questions not depending on conversation history answered by the answer field>
- deixis: <deixis applied to original variants>
- disfluencies: <disfluencies applied to original variants>
- deixis_disfluencies: <disfluencies applied to deixis variants>
- answer: <always the object field from the turn triple, representing B's answer to any of the questions>
Ensure that the variants of the original have the subject variable (enclosed by []) as is and that the object is always the answer and is never part of the questions. Ensure there are exactly three variants of each type. All questions should mimic real world conversational questions.
USER: You have been provided with K triples (subject, predicate, object) from the knowledge graph corresponding directly to exact turns. The subject and object, in this case, are templates and enclosed by [], and the subject template should be used as is for questions in the original field. For example, for a triple ([person], gender, [x]), a question in the original field should always use the literal "[person]" without any deixis. The answer field should always be the turn's object template. Your task is to use this information to generate a coherent conversational question-answering session between A and B following the aforementioned template. Remember their roles exactly and ensure the conversation length is equal to the number of turns.
Examples:
Triples
Turn 1: ([cricketer], number of matches played/races/starts, [a])
⋮

Figure 3: The prompt used for Synthetic Question Template Generation in the *Voice* setting.

SYSTEM: You are an AI assistant tasked with generating a natural conversational question-answering session between two people, A and B, based on information from a knowledge graph, in the form of a list of triples. A will only ask questions, and they should be based on the subject type and predicate of each triple, while B will only answer with just the object and no extraneous information. To make the conversation more realistic, you should also include for A:
- deixis (words that refer to people, places, or things in the conversation history like this, their, that, it, they, them)
You only return JSON of the following form with key being an <int representing the turn number> mapping to:
- original: <list of three variants of standard single-turn questions not depending on conversation history answered by the answer field>
- deixis: <deixis applied to original variants>
- answer: <always the object field from the turn triple, representing B's answer to any of the questions>
Ensure that the variants of the original have the subject variable (enclosed by []) as is and that the object is always the answer and is never part of the questions. Ensure there are exactly three variants of each type. All questions should mimic real world user search queries and be short, lower case and never proper questions beginning with who/whom/what/when/which/how. Ensure to never generate proper questions for any variant of the four types of queries.
USER: You have been provided with K triples (subject, predicate, object) from the knowledge graph corresponding directly to exact turns. The subject and object, in this case, are templates and enclosed by [], and the subject template should be used as is for questions in the original field. For example, for a triple ([person], gender, [x]), a question in the original field should always use the literal "[person]" without any deixis. The answer field should always be the turn's object template. Your task is to use this information to generate a coherent conversational question-answering session between A and B following the aforementioned template. Remember their roles exactly and ensure the conversation length is equal to the number of turns.
Examples: We see in the following examples all variants take on user search query form and never start with one of a who, what, when, which, and how.
Triples
Turn 1: ([cricketer], number of matches played/races/starts, [a])
Turn 2: ([cricketer], date of birth, [b])
⋮

Figure 4: Prompt used for Synthetic Question Template Generation in the *Text* (Search) setting.

SYSTEM: You are an AI assistant tasked with generating a natural conversational question-answering session between two people, A and B, based on information from a knowledge graph, in the form of a list of tuples. A will only ask questions, and they should be based on the subject type, predicate, and relationship predicate of each tuple (potentially also an object from another tuple provided), while B will only answer with just the object and no extraneous information. To make the conversation more realistic, you should also include for A:
- deixis (words that refer to people, places, or things in the conversation history like this, their, that, it, they, them) applied to just the subject template (never to any of the objects included)
You only return JSON of the following form with key being an <int representing the turn number> mapping to:
- original: <list of three variants of standard single-turn questions not depending on conversation history answered by the answer field>
- deixis: <deixis applied to original variants>
- answer: <always the object field from the turn tuple, representing B's answer to any of the questions>
Ensure that the variants of the original have the subject variable (enclosed by []) as is and that the object is always the answer and is never part of the questions. Ensure there are exactly three variants of each type. All questions should mimic real world user search queries and be short, lower case and never proper questions beginning with who/whom/what/when/which/how. Ensure to never generate proper questions for any variant of the four types of queries.
USER: You have been provided with K tuples (subject, predicate, relationship_predicate, object) from the knowledge graph corresponding directly to exact turns. The subject and object, in this case, are templates and enclosed by [], and the subject template should be used as is for questions in the original field. For example, for a tuple ([person], marriage, related person, [a]), a question in the original field should always use the literal "[person]" without any deixis. You can also use the object field from any of the other tuples from the same predicate, if available, to craft better questions. The answer field should always be the turn's object template. Your task is to use this information to generate a coherent conversational question-answering session between A and B following the aforementioned template. Remember their roles exactly and ensure the conversation length is equal to the number of turns. Never use the object template corresponding to the turn ([a] in 1, [b] in 2, ...) in any of the turn's questions.
Examples: We see in the following examples all variants take on user search query form and never start with one of a who, what, when, which, and how.
Triples
Turn 1: ([movie], voice actor, performer, [a])
Turn 2: ([movie], voice actor, character, [b])
⋮

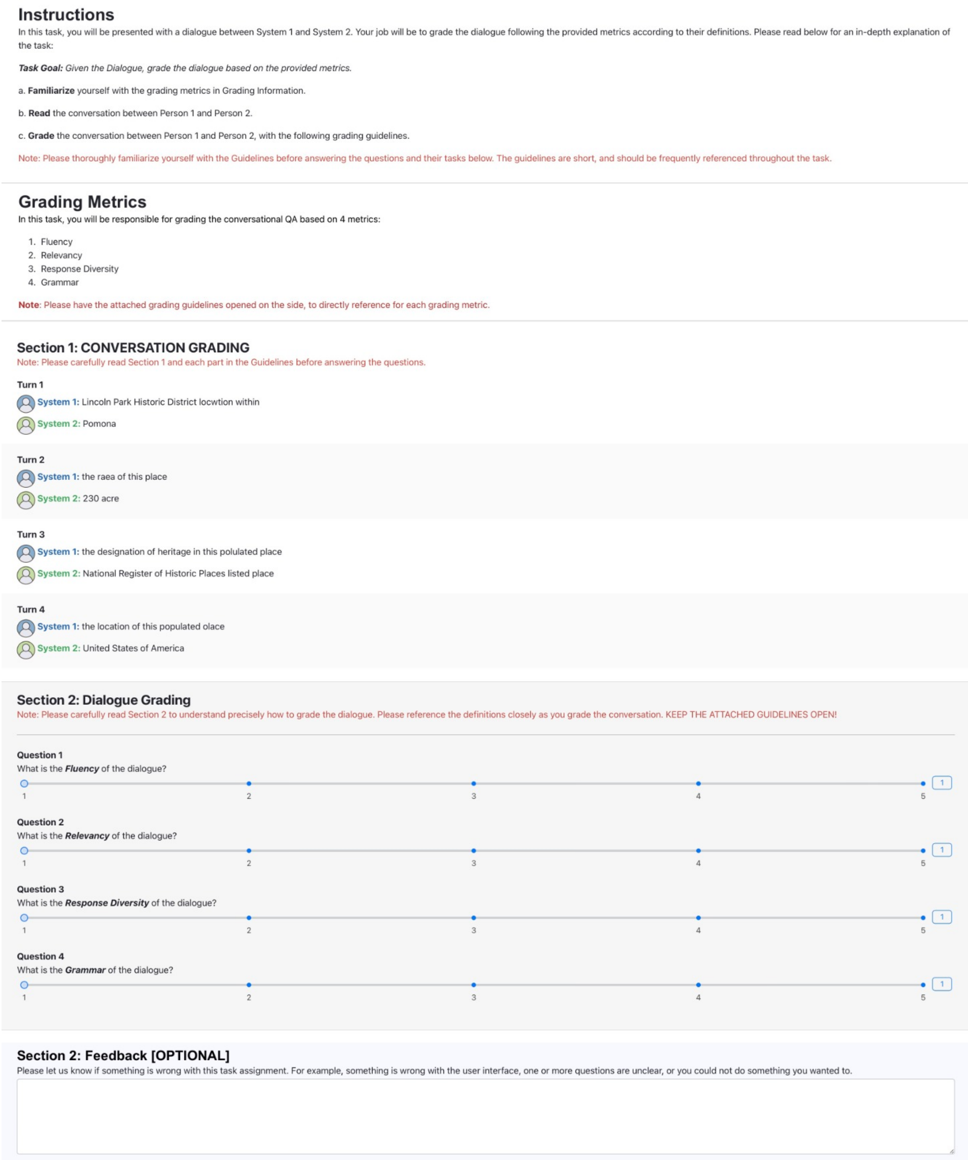
Figure 5: Prompt used for Synthetic Question Template Generation in the *Text* setting with Relationship Predicates.

SYSTEM: You are a helpful assistant that can do conversational factoid question answering. You only provide the exact answer span and never with extraneous information or in full sentences. Provide the answer in a string or pythonic list (the list can have multiple elements if there are multiple answers). Always provide an answer in the format "Answer: <answer string or list of answer strings>". If you are extremely unsure of the answer, return "Answer: NA".
USER: Who narrated the Penguins documentary?
GPT$_4$: *Ed Helms*
USER: Ummm, who was, hmm, its director?
GPT$_4$: *Alastair Fothergill*

Figure 6: Example Interaction for GPT$_\text{x}$ baselines of ConvKGYarn.

SYSTEM: You are a helpful assistant that can help evaluate conversational factoid question answering. You will be provided Questions, Gold Answers, and Candidates, turn-by-turn. The Gold Answer and Candidate are either a single answer or list of answers. If the Candidate seems to properly answer the question based on the answers, score it a 1, else, a 0. Do not use any of your global knowledge. If they are lists, ensure that at least one of the Candidate is captured by the Gold Answers. Do not use any additional knowledge. The output should be of the form R̈atings: <pythonic list of 0s/1s>ẅhere the listś order corresponds exactly to the conversation turn
USER: Question: Who narrated the Penguins documentary?
Gold Answers: Ed Helms Candidates: Ed Helms
Question: Ummm, who was, hmm, its director?
Gold Answers: Alastair Fothergill Candidates: NA
Question: Who produced the documentary?
Gold Answers: [Alastair Fothergill, Keith Scholey, Roy Conli]
Candidates: Scholey
GPT$_4$: *[1, 0, 1]*
⋮

Figure 7: Prompt for GPT$_4$-eval of ConvKGYarn.



Instructions

In this task, you will be presented with a dialogue between System 1 and System 2. Your job will be to grade the dialogue following the provided metrics according to their definitions. Please read below for an in-depth explanation of the task:

Task Goal: *Given the Dialogue, grade the dialogue based on the provided metrics.*

a. **Familiarize** yourself with the grading metrics in Grading Information.

b. **Read** the conversation between Person 1 and Person 2.

c. **Grade** the conversation between Person 1 and Person 2, with the following grading guidelines.

Note: Please thoroughly familiarize yourself with the Guidelines before answering the questions and their tasks below. The guidelines are short, and should be frequently referenced throughout the task.

Grading Metrics

In this task, you will be responsible for grading the conversational QA based on 4 metrics:

1. Fluency
2. Relevancy
3. Response Diversity
4. Grammar

Note: Please have the attached grading guidelines opened on the side, to directly reference for each grading metric.

Section 1: CONVERSATION GRADING

Note: Please carefully read Section 1 and each part in the Guidelines before answering the questions.

Turn 1

System 1: Lincoln Park Historic District locwtion within

System 2: Pomona

Turn 2

System 1: the raea of this place

System 2: 230 acre

Turn 3

System 1: the designation of heritage in this polulated place

System 2: National Register of Historic Places listed place

Turn 4

System 1: the location of this populated olace

System 2: United States of America

Section 2: Dialogue Grading

Note: Please carefully read Section 2 to understand precisely how to grade the dialogue. Please reference the definitions closely as you grade the conversation. KEEP THE ATTACHED GUIDELINES OPEN!

Question 1

What is the ***Fluency*** of the dialogue?

1 2 3 4 5 [1]

Question 2

What is the ***Relevancy*** of the dialogue?

1 2 3 4 5 [1]

Question 3

What is the ***Response Diversity*** of the dialogue?

1 2 3 4 5 [1]

Question 4

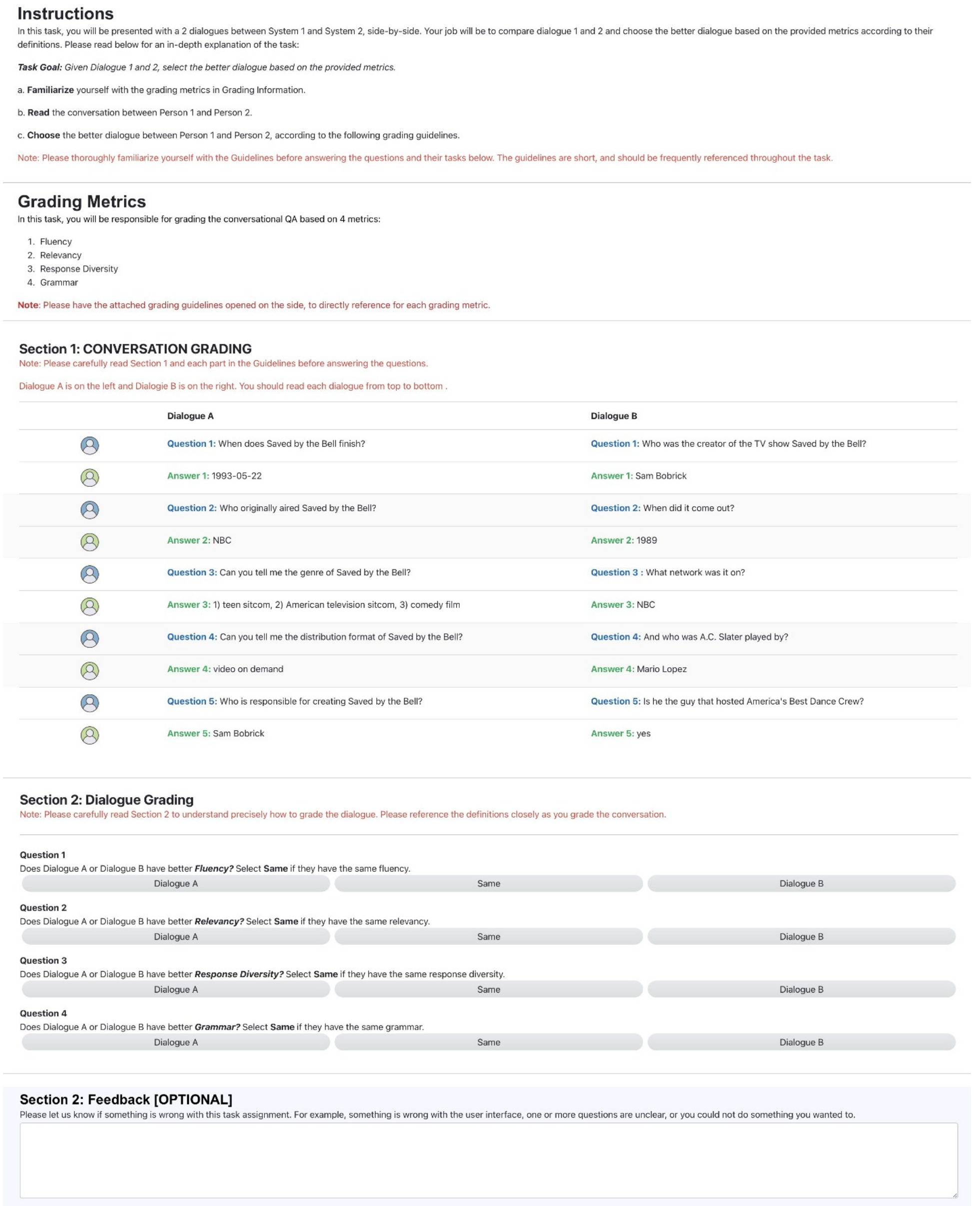
What is the ***Grammar*** of the dialogue?

1 2 3 4 5 [1]

Section 2: Feedback [OPTIONAL]

Please let us know if something is wrong with this task assignment. For example, something is wrong with the user interface, one or more questions are unclear, or you could not do something you wanted to.

Figure 8: The human annotation user interface for the Psychometric Evaluation of ConvKGYarn.



Instructions

In this task, you will be presented with a 2 dialogues between System 1 and System 2, side-by-side. Your job will be to compare dialogue 1 and 2 and choose the better dialogue based on the provided metrics according to their definitions. Please read below for an in-depth explanation of the task:

Task Goal: *Given Dialogue 1 and 2, select the better dialogue based on the provided metrics.*

a. **Familiarize** yourself with the grading metrics in Grading Information.

b. **Read** the conversation between Person 1 and Person 2.

c. **Choose** the better dialogue between Person 1 and Person 2, according to the following grading guidelines.

Note: Please thoroughly familiarize yourself with the Guidelines before answering the questions and their tasks below. The guidelines are short, and should be frequently referenced throughout the task.

Grading Metrics

In this task, you will be responsible for grading the conversational QA based on 4 metrics:

1. Fluency
2. Relevancy
3. Response Diversity
4. Grammar

Note: Please have the attached grading guidelines opened on the side, to directly reference for each grading metric.

Section 1: CONVERSATION GRADING

Note: Please carefully read Section 1 and each part in the Guidelines before answering the questions.

Dialogue A is on the left and Dialogie B is on the right. You should read each dialogue from top to bottom .

	Dialogue A	Dialogue B
	Question 1: When does Saved by the Bell finish?	**Question 1:** Who was the creator of the TV show Saved by the Bell?
	Answer 1: 1993-05-22	**Answer 1:** Sam Bobrick
	Question 2: Who originally aired Saved by the Bell?	**Question 2:** When did it come out?
	Answer 2: NBC	**Answer 2:** 1989
	Question 3: Can you tell me the genre of Saved by the Bell?	**Question 3 :** What network was it on?
	Answer 3: 1) teen sitcom, 2) American television sitcom, 3) comedy film	**Answer 3:** NBC
	Question 4: Can you tell me the distribution format of Saved by the Bell?	**Question 4:** And who was A.C. Slater played by?
	Answer 4: video on demand	**Answer 4:** Mario Lopez
	Question 5: Who is responsible for creating Saved by the Bell?	**Question 5:** Is he the guy that hosted America's Best Dance Crew?
	Answer 5: Sam Bobrick	**Answer 5:** yes

Section 2: Dialogue Grading

Note: Please carefully read Section 2 to understand precisely how to grade the dialogue. Please reference the definitions closely as you grade the conversation.

Question 1
Does Dialogue A or Dialogue B have better ***Fluency?*** Select **Same** if they have the same fluency.
Dialogue A | Same | Dialogue B

Question 2
Does Dialogue A or Dialogue B have better ***Relevancy?*** Select **Same** if they have the same relevancy.
Dialogue A | Same | Dialogue B

Question 3
Does Dialogue A or Dialogue B have better ***Response Diversity?*** Select **Same** if they have the same response diversity.
Dialogue A | Same | Dialogue B

Question 4
Does Dialogue A or Dialogue B have better ***Grammar?*** Select **Same** if they have the same grammar.
Dialogue A | Same | Dialogue B

Section 2: Feedback [OPTIONAL]

Please let us know if something is wrong with this task assignment. For example, something is wrong with the user interface, one or more questions are unclear, or you could not do something you wanted to.

Figure 9: The human annotation user interface for the Psychometric Comparative Analysis of ConvKGYarn.

Dialogue Grading - Task Guidelines

INTRODUCTION

Goal: The goal of this task is to grade the conversational QA, based on the provided metrics. Provided below is background information that will be useful for better understanding the task:

- **What is a Conversational QA?** Conversational QA means a conversation between two systems, that requests information at each turn. An example of this could be:

    System 1: How old is Ryan Reynolds?
    System 2: 46 years old

    System 1: What is Ryan Reynold's next movie?
    System 2: Deadpool 3

    System 1: When does Deadpool 3 come out?
    System 2: May 3, 2024

  You could interpret it as a Q&A session between two people.

- **What is a TURN?** A turn in the conversation is a round of a conversation. Essentially, once Person 1 and Person 2 speak once each. An example is highlighted in its turns:

    Turn 1
    System 1: How old is Ryan Reynolds?
    System 2: 46 years old

    Turn 2
    System 1: What is Ryan Reynold's next movie?
    System 2: Deadpool 3

    Turn 3
    System 1: When does Deadpool 3 come out?
    System 2: May 3, 2024

  Each highlight color, is a different turn.

TASK OVERVIEW

In this task, you will be presented with a Conversational QA between 2 systems. Your job will be to:

1. Read through the conversation, and understand each question and answer.
2. Thoroughly understand the grading metrics, and the examples for each.
3. Grade the conversation for each of the metrics.

Please ensure you read Section 1 of the guidelines before you grade the conversations.

GRADING METRICS

In this task, you will be responsible for grading the conversational QA based on 4 metrics:

1. Fluency
2. Relevancy
3. Response Diversity
4. Grammar

Please read below for a thorough understanding of each grading metric.

FLUENCY

DEFINITION

Fluency refers to the degree to which the content reads with ease, resembling natural human language. Fluent text will flow smoothly, sound authentic, and avoid awkward phrasings or constructions that might indicate machine generation or a non-native speaker.

In short, it is the ease and naturalness with which the text conveys information.

TIPS

Provided below are some tips in evaluating the fluency of the text:

- **How well does the text flow?**
    - Read the conversation out loud. This will help you identify any awkward or unnatural-sounding phrases.
- **How is the sentence structure?**
    - Sentences should be structured in a logical and well-read way, and should flow well. It should not sound choppy.
- **How is the vocabulary?**
    - The use of appropriate vocabulary can impact fluency.
    - Words used should be natural to the target text. If the style and terminology of the text is not appropriate, it is not fluent.
- **Stay Objective:**
    - Remember, fluency grading is about the flow of language, not the accuracy of content or the validity of ideas. Keep personal biases and content preferences separate from your fluency assessment.

GRADING SCALE

Note: You are only grading the Fluency of the conversation. You should not grade the content of the conversation or grammar.

To assess the fluency of the conversational QA, please read below:

Grading Level	Definition	Example of Levels of Fluent Text
1 - Basic	The text reads awkwardly and is often stilted or disjointed. The phrasing feels forced or unnatural, making it evident that the content might not have been written by a native speaker or is machine-generated. Text is basic, often fragmented, and may miss key connecting words.	**Translated Question:** "Biggest mountain what?" **Reason:** Technically, the meaning of the question is there. However, the text is awkward, and does not read well. There are fragments of information not a cohesive sentence. In addition, "biggest" would not be commonly be used to ask about the tallest mountain.
2 - Elementary	While the primary message of the text is decipherable, it still contains noticeable unnatural phrasings. The flow is better than the beginner level but requires the reader to make some effort to interpret the intended meaning. The text is more structured than the beginner level but might still lack proper phrasing.	**Translated Question:** "What mountain biggest?" **Reason**: The structure of the sentence is slightly better. At least the ordering is correct, in terms of asking for the information you're looking for, about the entity.
3 - Limited	The text reads more naturally with occasional lapses in fluency. Most of the content flows logically and sounds human-like, with only sporadic awkward phrasings or vocabulary choices. The text is clearer, conveying straightforward information with better structure. Word choices are more natural as well.	**Translated Question:** "What is the mountain with the maximum elevation on Earth?" **Reason:** This translation is technically correct. It has the correct structure, and gets the point across. It almost sounds, robotic, due to its technical nature. However, it sounds artificial, using technically correct language, that wouldn't be commonly used. More common variants are "tallest" or the "highest".
4 - Professional	The text closely resembles natural human language, with varied and appropriate phrasings. While it is coherent and mostly fluid, keen readers might spot occasional hints of non-human or non-native origins. Text is well-structured and clear, with a slight depth that adds context without adding complexity. Words are largely well chosen; however, may not be what a native speaker may choose.	**Translated Question:** "Which is the tallest mountain in the world?" **Reason:** This would be a perfectly fine way to phrase the source question. However, there is only one disrepancy, that differs from truly natural and local translations. Instead of "which", most people would use "what".
5 - Native	The text reads effortlessly, with the elegance and nuance of a seasoned human writer. It feels entirely authentic, with a rhythm and tone that aligns with natural human communication, leaving no traces of artificiality. Text is straightforward, fully natural, and effortlessly conveys the intended information or question. Words choices are native as well.	**Translated Question:** "What is the tallest mountain in the world?" **Reason:** This is a perfect question, of what the tallest mountain in the world is. The sentence structure is correct, and is how native people would ask the question.

YOUR JOB IS TO ONLY GRADE THE CONVERSATION FOR FLUENCY. IN ADDITION, DO NOT DOCK MARKS FOR GRAMMAR (SPELLING, PUNCTUATION, CAPITALIZATION) ERRORS UNLESS IT SIGNIFICANTLY IMPACTS FLUENCY.

RELEVANCY

DEFINITION

Relevancy in a conversation is measured by the extent to which each turn or statement is related to the preceding one. A conversation with high relevancy should maintain a consistent topic or theme, evolving organically without abrupt or unrelated deviations. Conversations that drift into unrelated subjects with little or no connection display lower relevancy.

TIPS

Provided below are some tips in evaluating the relevancy of the conversation:

- **Clearly Understand the Definition:**
    - Before grading, ensure that you fully comprehend what "relevancy" means in the context of a conversation. It refers to how connected or related consecutive statements or questions are to each other.
- **Listen or Read Actively:**
    - Pay close attention to the entire conversation, making mental or physical notes about where the conversation might drift from the topic.
- **Identify the Central Topic:**
    - Try to pinpoint the main topic or theme of the conversation. This serves as your reference point for determining how other parts of the conversation relate back to it.
- **Check for Natural Transitions:**
    - A conversation can evolve, but if it does so, there should be a natural and understandable transition from one topic to the next. If a topic shift feels abrupt or forced, it might indicate lower relevancy.
- **Avoid Personal Bias:**
    - Ensure that personal knowledge or feelings about the topic don't influence your grading. What might seem irrelevant to one person might be highly pertinent to another based on their experiences or knowledge base.

GRADING SCALE

Grading Level	Definition	Example of Levels of Relevant Text
1 - Not Relevant	Turns in the conversation have no clear connection to each other. The conversation jumps between unrelated topics with no transition.	**Turn 1**: "How old is Leonardo DiCaprio?" **Turn 2:** "How many moons does Jupiter have?" **Turn 3**: "When was the Eiffel Tower completed?" **Turn 4**: "What is the boiling point of water?" **Reason**: None of these questions correlate with each other on the theme or information.
2- Slightly Relevant	Some attempts at connection between topics, but many turns in the conversation feel forced or out of place.	**Turn 1**: "Which movie did Steven Spielberg direct in 1993?" **Turn 2**: "Who composed the music for 'The Dark Knight'?" **Turn 3**: "How old is Queen Elizabeth II?" **Turn 4**: "Who was the first president of the United States?" **Reason**: The questions are not well connected. However, there is an overarching concepts connecting them. Turn 1 and 2 has "movies" and Turn 3 and 4 have "political figures". There is an attempt to connect the questions; however, does not feel natural.
3 - Moderately Relevant	Most turns in the conversation relate to a central topic, but there are occasional drifts into unrelated subjects.	**Turn 1:** "What's the height of Mount Everest?" **Turn 2:** "Where is K2, the second-highest mountain, located?" **Turn 3**: "Who starred as the Joker in the 2008 film 'The Dark Knight'?" **Turn 4**: "In which Batman film did Arnold Schwarzenegger play the role of Mr. Freeze?" **Reason:** Some of the turns directly correlate with each other, but the entire conversation is not fluid. Turn 2 to Turn 3 does not make sense how the connection was made.
4 - Highly Relevant	Nearly all turns in the conversation have clear ties to a main topic or theme, with minimal deviation.	**Turn 1**: "When did World War II start?" **Turn 2**: "Which countries were part of the Axis Powers during World War II?" **Turn 3**: "When was Canada founded?" **Turn 4**: "Who were the first settlers in Canada?" **Reason:** Technically, each turn in the conversation has the connection to the next. However, the connections do not seem too natural in a conversation.
5 - Completely Relevant	Every turn in the conversation seamlessly flows from one to the next, maintaining a single, clear focus throughout.	**Turn 1**: "How many novels did Jane Austen write?" **Turn 2**: "Which of Jane Austen's novels was published while she was alive?" **Turn 3**: "Which year was Pride and Prejudice published?" **Turn 4**: "Who is the main character in Pride and Prejudice'?" **Reason:** Each turn in the conversation relate to each other, and the entire conversation has a central theme and intuitive flow.

RESPONSE DIVERSITY

DEFINITION

Response Diversity assesses the breadth and variety of questions posed within a conversation. A conversation with high response diversity will exhibit a broad spectrum of question types related to different entities, ensuring the conversation isn’t limited to a single topic or entity. The conversation should intuitively transition between topics while maintaining coherence and context.

TIPS

Provided below are some tips in evaluating the fluency of the text:

- **Contextual Comprehension**:
  - While diversity is crucial, it should not come at the expense of the conversation's coherence or relevance. A diverse conversation should still make logical sense. It's essential to evaluate how smoothly and intuitively topics transition from one to another. A conversation that jumps between entirely unrelated entities without a connecting thread may be diverse but can be perceived as disjointed or lacking depth.
- **Depth vs. Breadth**:
  - Diversity isn't just about the quantity of topics or entities touched upon; it's also about the depth with which each topic is explored. A conversation that skims the surface of ten topics may be less valuable than one that dives deeply into three and effectively links them. When grading, consider a balance between depth (how comprehensively each topic is covered) and breadth (how many different topics or entities are introduced).
- **Variability in Question Types**:
  - Diversity also involves varying the kind of questions posed. For instance, a conversation that includes a multiple aspects of an entity (ex. age, height, birthdate) has richer diversity vs. asking about one topic (ex. age only).

Remember, the goal of grading response diversity is to encourage a multifaceted, enriching, and engaging conversation that covers a broad spectrum without losing focus or coherence.

GRADING SCALE

Grading Level	Definition	Example of Levels of Relevant Text
1 - Low Diversity	Questions predominantly focus on a single entity or topic, with minimal or no variation in the type of questions asked.	**Example:** "What is the Mona Lisa? Who painted the Mona Lisa? When was the Mona Lisa painted? What's the history of the Mona Lisa?" **Reason:** Only asking surface level questions about Mona Lisa.
2- Below Average Diversity	Shows slight variation in entities or topics, but the types of questions remain largely consistent or predictable.	**Example**: "What is the Mona Lisa? Who painted the Mona Lisa? Who painted The Last Supper? When was The Starry Night painted?" **Reason**: Has more diversity in the type of questions asked, and traverses different entities. Goes from Mona Lisa, to The Last Supper and The Starry Night. But is stuck on Leonardo DaVinci-related content. As well, "who painted" and "when was" questions.
3 - Moderately Diversity	Displays a mix of different entities or topics with some variety in question types, but might lack a smooth transition or coherence between them.	**Example:** What is the Mona Lisa? Who was the most famous painter in the Renaissance era? What is the most expensive art piece from the Renaissance era? **Reason:** Has more diversity of of entities and and the type of questions that are asked across the different entities themselves. But it is stuck in the smaller realm of art.
4 - Above Average Diversity	Broad range of question types covering multiple entities or topics with coherent transitions, but may occasionally revert to a specific topic or exhibit minor lapses.	**Example**: "What is the Mona Lisa? Leonardo Da Vinci's famous artworks? Other influential art figures in the Renaissance era? **Reason:** Although the question type changes, and the entities switch, it's only in the scope of art in the Renaissance era. That said, it is a broader scope, and there is more exploration across entities and topics.
5 - High Diversity	Demonstrates a wide spectrum of question types related to various entities, with seamless transitions and consistent coherence throughout the conversation.	**Example:** "What is the Mona Lisa? Famous Rennessaince painters in Europe? Is Beetoven Renaissance music? Can you name some contemporary artists inspired by classical art?" **Reason:** It traverses various entities, and asks unique questions about each of them, while still in the bounds of logical flow.

GRAMMAR

Definition: Grammatical correctness refers to the adherence to established rules and conventions of a particular language regarding sentence structure, verb conjugation, punctuation, word order, and other syntactic and morphological elements. It ensures clarity, consistency, and proper communication within that language. However, it's essential to recognize that these rules can vary significantly between languages, and what's deemed grammatically correct in one language might not be in another.

Grammar focuses on the technical correctness of language. This is different from fluency which emphasizes the flow, ease, and naturalness of communication. Grammar refers to the system and structure of a language, emphasizing the proper arrangement of words and phrases to create well-formed sentences. It's about the rules and technical aspects of a language.

TIPS

Provided below are some tips in evaluating the fluency of the text:

- Familiarize with Language Specifics:
  - Before grading, understand English grammar rules.
- Review Basic Elements:
  - Check for subject-verb agreement, proper tense usage, and correct word order.
- Evaluate Punctuation:
  - Ensure the correct usage of commas, periods, semicolons, and other punctuation marks relevant to the specific language.
- Check Sentence Structures:
  - Ensure variety in sentence types (e.g., declarative, interrogative) and look for sentence fragments or run-ons.
- Assess Word Choice:
  - Verify the correct usage of homonyms, synonyms, and other language-specific intricacies.
- Examine Modifiers:
  - Ensure modifiers (like adjectives and adverbs) are placed correctly and aren't dangling or misplaced.

Remember to stay objective. Different languages have unique rules. Don't impose the conventions of one language onto another.

GRADING SCALE

Note: You are only grading the Grammar of the translated text. You should not grade the content of the conversation.

To assess the grammar of the Translated Question, please read below:

Grading Level	Definition	Examples of Levels of Grammar
1 - Beginner	Contains fragmented sentences and numerous grammatical errors that greatly affect comprehension. Has many spelling errors.	**Translated Question:** "eiffel tower were is?" **Reason:** The overall structure of the sentence is incorrect. In addtion, Eiffel Tower was not capitlized. "Where" is not correctly spelled. Due to the errors, the question might not be understandable.
2 - Novice	Has multiple grammatical mistakes but the central question or point is discernible. Has some spelling errors.	**Translated Question:** "Where Eiffel Tower located." **Reason:** The overall sentence structure is better; however there are several missing words "Where **is"** and a question mark is not used at the end of the question. You can understand the question, but it is obvious there are mistakes.
3 - Intermediate	Displays occasional grammatical errors but the message remains clear. Has only a couple spelling errors.	**Translated Question:** "Where are the Eifel Tower location?" **Reason:** Eiffel Tower is incorrectly spelled, and "where are" should be "where is" due to it being singular. You can easily understand the question; however, there are a couple minor errors.
4 - Advanced	Demonstrates very few and minor grammatical errors that don't hinder comprehension. Potentially has a single spelling error.	**Translated Question:** "Where does the Eiffel Tower located?" **Reason:** The sentence structure is correct, but there is a minor mistake of using "does" instead of "is".
5 - Expert	Showcases exemplary grammar without errors.	**Translated Question:** "Where is the Eiffel Tower located?" **Reason:** The translated question has correct grammar, consisting of correct sentence structure, punctuation and capilization

Note: Grade 1, is largely about major mistakes that can inhibit understanding. Grades 2-4, are largely about the quantity of errors. Grade 5, is perfect grammar.

Dialogue Comparisons - Task Guidelines

INTRODUCTION

Goal: The goal of this task is to compare two system dialogues, based on the provided metrics. Provided below is background information that will be useful for better understanding the task:

- **What is a Conversational QA?** Conversational QA means a conversation between two systems, that requests information at each turn. An example of this could be:

System 1: How old is Ryan Reynolds?
System 2: 46 years old

System 1: What is Ryan Reynold's next movie?
System 2: Deadpool 3

System 1: When does Deadpool 3 come out?
System 2: May 3, 2024

You could interpret it as a Q&A session between two people.

- **What is a TURN?** A turn in the conversation is a round of a conversation. Essentially, once Person 1 and Person 2 speak once each. An example is highlighted in its turns:

Turn 1
System 1: How old is Ryan Reynolds?
System 2: 46 years old

Turn 2
System 1: What is Ryan Reynold's next movie?
System 2: Deadpool 3

Turn 3
System 1: When does Deadpool 3 come out?
System 2: May 3, 2024

Each highlight color, is a different turn.

TASK OVERVIEW

In this task, you will be presented with a Conversational QA between 2 systems. Your job will be to:

1. Read through the conversation, and understand each question and answer.
2. Thoroughly understand the grading metrics, and the examples for each.
3. Choose which dialogue is better, or if they are the same, for the given grading metric.

Please ensure you read Section 1 of the guidelines before you compare the dialogues.

GRADING METRICS

In this task, you will be responsible for comparing the conversational QA dialogues based on 4 metrics:

1. Fluency
2. Relevancy
3. Response Diversity
4. Grammar

Please read below for a thorough understanding of each grading metric.

FLUENCY

DEFINITION

Fluency refers to the degree to which the content reads with ease, resembling natural human language. Fluent text will flow smoothly, sound authentic, and avoid awkward phrasings or constructions that might indicate machine generation or a non-native speaker.

In short, it is the ease and naturalness with which the text conveys information.

TIPS

Provided below are some tips in evaluating the fluency of the text:

- **How well does the text flow?**
  - Read the conversation out loud. This will help you identify any awkward or unnatural-sounding phrases.
- **How is the sentence structure?**

  - Sentences should be structured in a logical and well-read way, and should flow well. It should not sound choppy.
- **How is the vocabulary?**
  - The use of appropriate vocabulary can impact fluency.
  - Words used should be natural to the target text. If the style and terminology of the text is not appropriate, it is not fluent.
- **Stay Objective:**
  - Remember, fluency grading is about the flow of language, not the accuracy of content or the validity of ideas. Keep personal biases and content preferences separate from your fluency assessment.

RELEVANCY

DEFINITION

Relevancy in a conversation is measured by the extent to which each turn or statement is related to the preceding one. A conversation with high relevancy should maintain a consistent topic or theme, evolving organically without abrupt or unrelated deviations. Conversations that drift into unrelated subjects with little or no connection display lower relevancy.

TIPS

Provided below are some tips in evaluating the relevancy of the conversation:

- **Clearly Understand the Definition:**
  - Before grading, ensure that you fully comprehend what "relevancy" means in the context of a conversation. It refers to how connected or related consecutive statements or questions are to each other.
- **Listen or Read Actively:**
  - Pay close attention to the entire conversation, making mental or physical notes about where the conversation might drift from the topic.
- **Identify the Central Topic:**
  - Try to pinpoint the main topic or theme of the conversation. This serves as your reference point for determining how other parts of the conversation relate back to it.
- **Check for Natural Transitions:**
  - A conversation can evolve, but if it does so, there should be a natural and understandable transition from one topic to the next. If a topic shift feels abrupt or forced, it might indicate lower relevancy.
- **Avoid Personal Bias:**
  - Ensure that personal knowledge or feelings about the topic don't influence your grading. What might seem irrelevant to one person might be highly pertinent to another based on their experiences or knowledge base.

RESPONSE DIVERSITY

DEFINITION

Response Diversity assesses the breadth and variety of questions posed within a conversation. A conversation with high response diversity will exhibit a broad spectrum of question types related to different entities, ensuring the conversation isn't limited to a single topic or entity. The conversation should intuitively transition between topics while maintaining coherence and context.

TIPS

Provided below are some tips in evaluating the fluency of the text:

- **Contextual Comprehension**:
  - While diversity is crucial, it should not come at the expense of the conversation's coherence or relevance. A diverse conversation should still make logical sense. It's essential to evaluate how smoothly and intuitively topics transition from one to another. A conversation that jumps between entirely unrelated entities without a connecting thread may be diverse but can be perceived as disjointed or lacking depth.
- **Depth vs. Breadth**:
  - Diversity isn't just about the quantity of topics or entities touched upon; it's also about the depth with which each topic is explored. A conversation that skims the surface of ten topics may be less valuable than one that dives deeply into three and effectively links them. When grading, consider a balance between depth (how comprehensively each topic is covered) and breadth (how many different topics or entities are introduced).
- **Variability in Question Types**:
  - Diversity also involves varying the kind of questions posed. For instance, a conversation that includes a multiple aspects of an entity (ex. age, height, birthdate) has richer diversity vs. asking about one topic (ex. age only).

Remember, the goal of grading response diversity is to encourage a multifaceted, enriching, and engaging conversation that covers a broad spectrum without losing focus or coherence.

GRAMMAR

DEFINITION

Grammatical correctness refers to the adherence to established rules and conventions of a particular language regarding sentence structure, verb conjugation, punctuation, word order, and other syntactic and morphological elements. It ensures clarity, consistency, and proper communication within that language. However, it's essential to recognize that these rules can vary significantly between languages, and what's deemed grammatically correct in one language might not be in another.

Grammar focuses on the technical correctness of language. This is different from fluency which emphasizes the flow, ease, and naturalness of communication. Grammar refers to the system and structure of a language, emphasizing the proper arrangement of words and phrases to create well-formed sentences. It's about the rules and technical aspects of a language.

TIPS

Provided below are some tips in evaluating the fluency of the text:

- Familiarize with Language Specifics:
    - Before grading, understand English grammar rules.
- Review Basic Elements:
    - Check for subject-verb agreement, proper tense usage, and correct word order.
- Evaluate Punctuation:
    - Ensure the correct usage of commas, periods, semicolons, and other punctuation marks relevant to the specific language.
- Check Sentence Structures:
    - Ensure variety in sentence types (e.g., declarative, interrogative) and look for sentence fragments or run-ons.
- Assess Word Choice:
    - Verify the correct usage of homonyms, synonyms, and other language-specific intricacies.
- Examine Modifiers:
    - Ensure modifiers (like adjectives and adverbs) are placed correctly and aren't dangling or misplaced.

Remember to stay objective. Different languages have unique rules. Don't impose the conventions of one language onto another.