

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

Ronak Pradeep^{1,2*}, Daniel Lee^{1,3*}, Ali Mousavi¹, Jeff Pound¹, Yisi Sang¹, Jimmy Lin², Ihab Ilyas¹, Saloni Potdar¹, Mostafa Arefiyan¹ and Yunyao Li^{3*}

¹ Apple ² University of Waterloo ³ Adobe

rpradeep@uwaterloo.ca {mostafaa, s_potdar}@apple.com
yunyao1@adobe.com

Abstract

The rapid evolution 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 demonstrate ConvKGYarn’s effectiveness in producing high-quality data comparable to popular conversational KGQA datasets across various metrics. ConvKGYarn excels in adhering to human interaction configurations and operating at a significantly larger scale. We showcase ConvKGYarn’s 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 ConvKGYarn 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 LLMs and conversational assistants in daily user interactions comes with the need for dynamic datasets to stress-test their ability to handle evolving knowledge-seeking questions. KGs have long been recognized for capturing structured representations of the world (Hogan et al., 2021). They represent concepts and entities as nodes, while edges form semantic relationships to

define facts. KGs have strong roots in various fields, including Natural Language Processing (Schneider et al., 2022), Recommender Systems (Guo et al., 2022), and Information Retrieval (Reinanda et al., 2020).

Integrating LLMs with KGs has advanced several NLP tasks (Petroni et al., 2019; Guu et al., 2020; Barba et al., 2021; Chakrabarti et al., 2022; Xu et al., 2023). This synergy unlocks new avenues for conversational KGQA scenarios like those targeted by ConvQuestions (Christmann et al., 2019). ConvQuestions highlights the potential of combining LLMs with KGs for accurate and attributed responses in conversations (Christmann et al., 2023).

Recent advancements in text retrieval have demonstrated the efficacy of LLM-generated synthetic data in enhancing downstream systems, from query synthesis (Nogueira and Lin, 2019; Ma et al., 2022; Pradeep et al., 2022) and LLM-based ranked list reorderings (Pradeep et al., 2023a,b; Tamber et al., 2023) to training highly effective small-scale models through automated prompt optimization (Xian et al., 2024). These developments underscore the opportunity to leverage synthetic data strategies from LLMs.

However, existing QA datasets lag behind evolving user needs. We introduce ConvKGYarn, a method for generating large-scale configurable conversational KGQA datasets. Psychometric evaluation show ConvKGYarn produces high-quality conversational data, scaling entity and fact coverage while incorporating diverse user interaction styles.

Evaluating ConvKGYarn-generated datasets with various LLMs reveals their struggle with fact recall, emphasizing the need for retrieval-augmented systems. Model effectiveness varies across different user interaction styles, highlighting the importance of building adaptable LLMs.

Through this work, we aim to shed light on building evolving datasets that can train and test conversational assistants of the future.

*Work done while at Apple

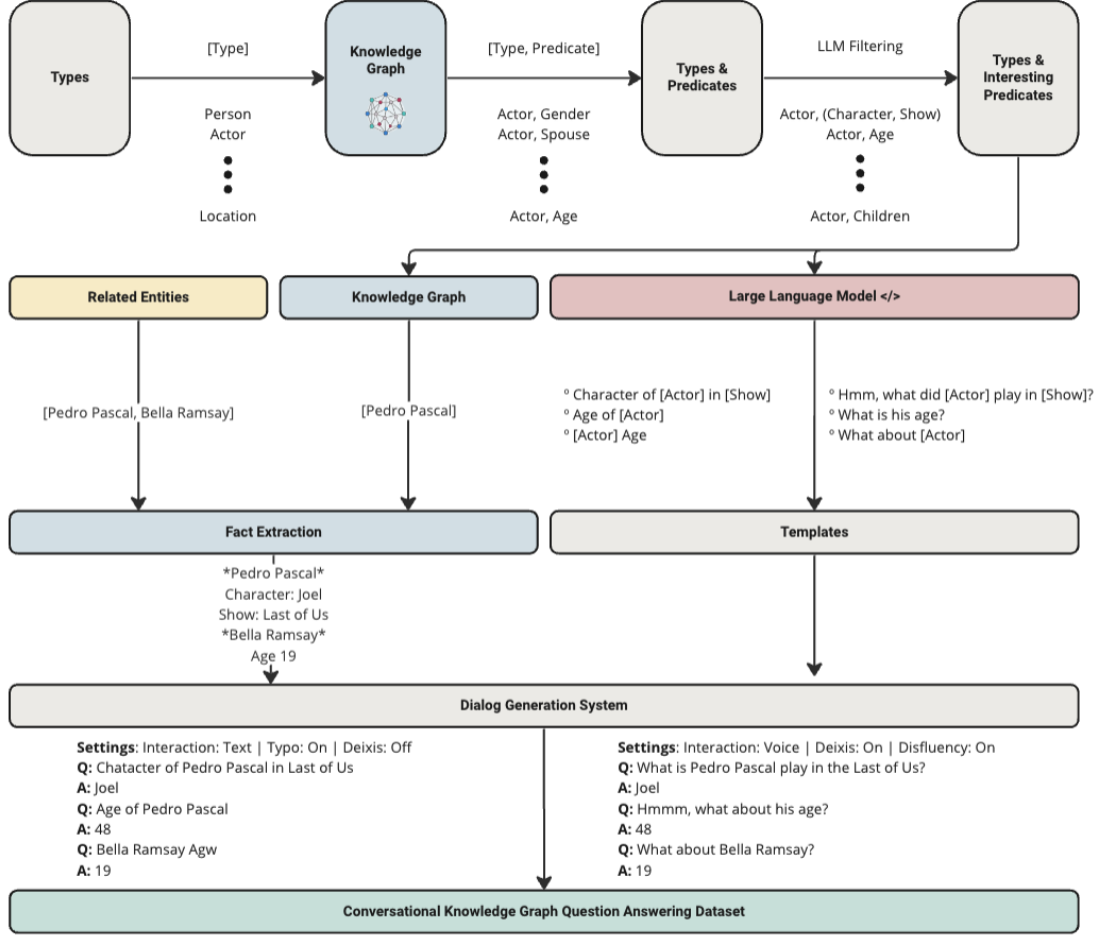


Figure 1: The full ConvKGYarn pipeline.

2 ConvKGYarn

Figure 1 illustrates the entire pipeline of the ConvKGYarn system. We first introduce the key notations and definitions to set our terminological and conceptual framework. Next, we dive into each module that comprises the ConvKGYarn system.

2.1 Definitions and Notations

The knowledge graph (KG) serves as our foundation. Following Wikidata terminology, an item (or entity) $e \in \mathcal{E}$ is described by statements (or facts) \mathcal{S}_e represented as *item-property-value* tuples. Properties (or predicates) are denoted by $p_e \in \mathcal{P}_e$.

Values (or objects) for a particular entity and predicate p_e are denoted by o_{p_e} . In ConvKGYarn, a *simple fact* refers to a property-predicate pair where the predicate does not involve multiple entries. Some entities possess properties with multiple values, such as an Actor’s siblings or a Country’s official languages; these are *complex facts* in ConvKGYarn. Additionally, *qualified facts* include different values with qualifiers (e.g., a Country’s population or a company’s CEO with timestamps).

These qualifiers refine the values within a statement and are supported in ConvKGYarn.

Each entity e is associated with multiple *types* T_e , with a specific type denoted by t_e (e.g., Singer, Movie). In addition to using the InstanceOf predicate to describe types, we use the Occupation predicate to add nuances to these types (for Person). This distinction helps identify the *interesting* predicates relevant to different types, such as Politician versus 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 $\mathcal{F}(t) = \{p_1, \dots, p_n\}$, where \mathcal{F} is the extraction function, t is a type, and $\{p_1, \dots, p_n\}$ is the set of predicates such that there exists some entity e of type t , for which p_i is a valid predicate.

2.3 LLM Predicate Selector

This step employs a large language model (LLM) to filter extracted predicates, selecting the most *interesting* ones for each type. The process is governed

by the prompt shown in Figure 2 in Appendix A and can be formulated as $\mathcal{G}(t, \{p_1, \dots, p_n\}) = \{p'_1, \dots, p'_m\}$, where \mathcal{G} denotes the selector function, selecting a subset $\{p'_1, \dots, p'_m\}$ from the initial predicate set.

By prompting with high specificity, we aim to select predicates that enhance the dataset’s richness while maintaining contextual appropriateness. Including the Wikidata identifier of the predicate helps clarify cases where the identifier name is ambiguous, leveraging LLMs familiarity with Wikidata.

Predicates that pass through this filter are expected to contribute meaningfully to discussions about the entity type. To ensure this, we prompt the model to exclude overly generic predicates, irrelevant noise, or mere identifiers, as these do not enhance a high-quality conversational QA dataset.

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 $\mathcal{I}(e) = \{(e, p'_1, o_1), \dots, (e, p'_m, o_m)\}$, 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 ensure configurability and scalability, while maintaining the tractability of ConvKGYarn, we

generate questions using a templated approach. We incorporate placeholders for entity type (e.g., [actor]), interesting predicates, and placeholder objects ([i]) in the prompt. Detailed prompts for generating questions for voice interactions and textual (or search) interactions are presented in Figure 3 and Figure 4 in Appendix A, respectively. The prompt for qualified facts is provided in Figure 7 (in Appendix A).

Designing ConvKGYarn involved emulating 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, highlighting differences in user experience. For text, we mimic search queries, emphasizing short keyword queries with successive follow-ups. They enable *deixis*, where questions refer to previously mentioned entities, enhancing continuity. Additionally, ConvKGYarn accounts for typographical errors (typos) in a post-processing step discussed in Section 3. In voice interactions, we aim to generate well-formed questions. The modality allows for conversations with *disfluencies*, mimicking natural speech imperfections such as *uh*, *um*, takebacks, apologies, thanks, or repetitions. We combine these aspects in the “*deixis_disfluencies*” variants to simulate human conversation intricacies, 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 variation in generated questions compared to querying the LLM multiple times, which is slower, more expensive, and less likely to yield diverse outputs. Generating all variants together helps ensure consistency, providing comprehensive data with a wide range of linguistic variations, which better evaluates the robustness of conversational QA systems or LLMs.

To speed up inference, we provide five triples. Note that the turn number does not indicate generating a question for that specific turn but serves as an index for both the JSON key and the object identifier. The JSON format in the prompt is crucial for systematic data parsing during the generation process, ensuring consistent question formatting and easy integration into our pipeline.

For qualified facts, we generalized standard triples to tuples with an additional relational predicate field. While turn-specific objects are disallowed in questions, objects from other turns within the same predicate help create more com-

plex queries. For example, the query “voice of [a] in [movie]” could correspond to turn 2 of the prompt with the answer “[b]”.

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 available), can be slot-filled using examples from the generated templates to get a conversation instance. This process adheres to specific rules: the first turn never involves any deixis, regardless of the interaction type or selected linguistic phenomena. Predicates are grouped to ensure cohesiveness and avoid unusual artifacts in the final conversations. For instance, questions about the date of birth or place of birth are likely to occur together rather than being separated by several facts.

This method integrates current factual data from the KG with synthetic templates, which are verifiable by humans, to form a factoid KGQA instance. Given that template-based generation and slot-filling are significantly more cost-effective than generating specific conversations for each new entity, ConvKGYarn allows us to efficiently curate large-scale, configurable datasets.

2.8 Resourcing and Cost

The cost structure of ConvKGYarn is designed for efficiency and scalability. The majority of LLM-related expenses are incurred upfront during template generation.

For the LLM Predicate Selector, costs are based on the number of unique types and their associated predicates, with each call using approximately 4096 tokens. In our experiments, this step cost less than 100 USD.

The synthetic question template generation, while more intensive due to multiple interaction types, and fact types (simple, complex, and qualified), leverages a more cost-effective model to manage expenses. On average, this involves about 14 calls per entity type costing us around 500 USD.

Importantly, after these initial investments, the cost of generating new conversations scales very efficiently. The template-based approach and slot-filling mechanism allow for the creation of large-scale, configurable datasets at a fraction of the cost of generating specific conversations for each new entity. This makes ConvKGYarn a highly cost-effective solution for producing extensive, verifiable factoid KGQA instances.

3 Experimental Setup

We use a Wikidata dump with a June 2023 knowledge cut-off for all our experiments. The dump, with roughly 100M entities, was filtered to include only English entity names and *interesting* types, resulting in 29M entities with 196M facts.

For the LLM predicate selector, we used the gpt-4-0613 endpoint. We query at most 50 predicates to avoid overwhelming the model, processing each type-predicate pair segment by segment. For predicates with linked qualifiers, we include the relationship predicate in the input, selecting those relevant for conversational factoid QA.

For synthetic question template generation, we use the gpt-3.5-turbo endpoint, providing two in-context examples per prompt to align generations with the expected template format. To handle textual interactions, we utilize the “logit_bias” field to penalize the model when it generates question words (wh-words or how), ensuring adherence to instructions and in-context examples.

For typo augmentation, we apply one of the following TextAttack attacks (Morris et al., 2020) at random to each question turn: WordSwapRandomCharacterDeletion(), WordSwapNeighboringCharacterSwap(), or WordSwapQWERTY(). Each question turn receives a single “meaningful” typo. We introduce a single “meaningful” typo to each question turn.

4 Dataset Statistics

The *General* set from ConvKGYarn comprises 29M entities and 196M facts from filtered Wikidata, excluding related entities. Each fact can generate 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 enables diverse conversation generation, providing a large-scale resource for training conversational agents and exposing language models to high-quality synthetic data. The dataset includes 274 unique types and 1252 unique predicates, enhancing the complexity and realism of factoid conversations. This scale and coverage surpass human-curated datasets like ConvQuestions (Christmann et al., 2019), which contain 11K real-user conversations averaging five questions each, limited to five primary entity types.

In contrast, the *Related* set focuses on popular Human-type entities, containing 210K entities and 6.1M facts. Despite its smaller scale, it offers a

Interaction	Deixis	Disfluency	Typo	Fluency	Relevance	Diversity	Grammar	Agreement
Voice	✗	✗	-	3.97 / 3.70	4.63 / 3.71	2.40 / 2.66	3.90 / 3.69	75.5 / 68.5
Voice	✗	✓	-	3.39 / 3.34	4.49 / 3.74	2.25 / 2.59	3.37 / 3.39	73.5 / 67.6
Voice	✓	✗	-	3.99 / 3.76	4.59 / 3.89	2.45 / 2.79	3.77 / 3.72	74.8 / 69.3
Voice	✓	✓	-	3.29 / 3.36	4.41 / 3.73	2.32 / 2.73	3.02 / 3.38	71.0 / 71.5
Text	✗	-	✗	2.83 / 2.57	4.41 / 3.38	2.19 / 2.58	2.95 / 2.75	70.8 / 66.3
Text	✗	-	✓	2.61 / 2.29	4.36 / 3.45	2.17 / 2.45	2.18 / 1.97	68.8 / 71.5
Text	✓	-	✗	2.84 / 2.48	4.36 / 3.33	2.29 / 2.54	2.83 / 2.73	67.1 / 70.8
Text	✓	-	✓	2.29 / 2.12	4.09 / 3.31	2.00 / 2.58	1.63 / 1.86	73.0 / 68.5

Table 1: The results from the Single Model Rating of the *General* (ConvKGYarn_G) and *Related* (ConvKGYarn_R) sets (scores separated by /) 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.

high density of interconnected information with an average of 54 questions per fact (an additional 30 from related entity-specific follow-up questions). This set includes 95 unique types and 265 unique predicates, providing a targeted dataset for detailed exploration and evaluation of conversational systems focused on human-centric entities.

5 Results

To evaluate ConvKGYarn’s efficacy, we employ three complementary methods: (1) Single-Model Rating, (2) Pairwise Comparison, and (3) Parametric Knowledge Evaluation of LLMs.

Single-Model Rating, using Likert scores, offers scalability but has limitations. It relies on absolute judgments, which can be less reliable than relative comparisons (Stewart et al., 2005) and lead to biases among annotators (Kulikov et al., 2019).

Pairwise Comparison mitigates these issues by facilitating relative judgments. However, it becomes less efficient when comparing multiple models, often requiring re-evaluation of baselines upon introducing new models (Stewart et al., 2005).

Lastly, we assess the effectiveness of LLMs on ConvKGYarn-generated conversational factoid QA datasets, examining their fact recall abilities through LLM-as-a-Judge evaluation. While this scales well and often correlates strongly with human annotations, it may suffer from self-enhancement bias, where LLMs favor their own generated answers (Zheng et al., 2023).

This multifaceted approach ensures a comprehensive evaluation of ConvKGYarn from both human and automated perspectives, leveraging each method’s strengths to offset others’ weaknesses.

5.1 Single-Model Rating

The Single-Model Rating task involves human annotators scoring multi-turn conversations on a 1-5

scale across four parameters: Fluency, Relevance, Diversity, and Grammar. We evaluated 1600 conversations sampled uniformly across 16 combinations of ConvKGYarn pipeline settings, including Interaction (Voice/Text), Deixis (On/Off), Disfluency (On/Off for Voice), Typo (On/Off for Text), and Related Entities (On/Off). The dataset covers diverse entities from Wikidata, spanning types such as Person, Actor, Singer, and Politician. The details of the task interface and the annotation guidelines are in Appendix B.

Table 1 presents parameter scores against setting configurations. We see that typographical errors negatively impact fluency and grammar, as expected. Using deixis improves fluency, given better conversation flow. Relevance and diversity remain largely unaffected by deixis, disfluencies, typos, and interaction settings as desired given the consistent fact set. However, related entities enhance diversity by incorporating connected concepts from the KG, seemingly at the cost of relevance.

We believe the optimal configuration uses voice interaction with deixis and related entities, minus disfluencies, mimicking natural human discourse. Despite evaluation subjectivity, annotator agreement averages 70.53%, indicating good consensus and evaluation reliability.

These findings underscore the importance of multi-dimensional evaluation when assessing synthetic conversational datasets. By exploring the impacts of various factors on key parameters, we gain nuanced insights into the strengths and limitations of ConvKGYarn, informing future refinements.

5.2 Pairwise Comparison

The Pairwise Comparison task presents human annotators with two conversations: one generated by ConvKGYarn and another from a widely used

Type	Fluency (%)	Relevance (%)	Diversity (%)	Grammar (%)
Preference	56.6	56.6	45.5	62.2
Agreement	86.6	82.2	84.6	89.0

Table 2: The results from the Pairwise Comparison. Preference dictates the percentage of graders who prefer ConvKGYarn. Agreement describes the percentage of conversations where 2 or more annotators agreed.

conversational KGQA dataset. Annotators indicate their preferences for the same psychometric evaluation metrics described in Section 5.1 for 500 conversations, focusing on voice interaction without disfluencies and related entities.

We chose ConvQuestions (Christmann et al., 2019) as the reference dataset due to its similarity to ConvKGYarn’s purpose and capabilities while being human-curated. To ensure a fair comparison and avoid confounders, we adapted the ConvKGYarn process outlined in Section 5.1 with three modifications: we restricted the types of entities to those available in the benchmark dataset; the entity referenced in the first turn of the reference conversation was used as the starting entity in the ConvKGYarn process; and the number of turns in both data sets were equalized.

Table 2 reveals notable patterns in our results. ConvKGYarn shows slight improvements in fluency and relevance over human-curated reference conversations, with a 56.6% preference rate. This advantage likely stems from ConvKGYarn’s methodology, which generates questions from a diverse knowledge base encompassing primary and related entities. In contrast, human-curated conversations depend on annotators’ research of given entities, potentially introducing higher variability. The marginal fluency advantage may be attributed to the standardized dialect and writing style of the LLM used in ConvKGYarn, compared to the inherent variance across human annotators. These findings suggest that ConvKGYarn’s systematic approach to conversation generation can produce results comparable to, and in some aspects slightly superior to, human-curated datasets.

Grammar emerges as a dimension where ConvKGYarn significantly outperforms, with a 62.2% preference. This superiority can be attributed to the grammatical proficiency of LLMs in structuring highly accurate sentences for the English language and locale. However, diversity proves challenging for ConvKGYarn, with only a 45.5% preference. This limitation likely stems from the structured

method of generating questions based on entity types and KG relationships, potentially constraining topic range compared to the more open-ended human curation process. Human annotators demonstrate strong consensus with an average 85.6% agreement. Their textual feedback provides valuable qualitative insights into ConvKGYarn’s perceived strengths and weaknesses, offering a deeper understanding of its effectiveness beyond quantitative metrics.

The human evaluation of ConvKGYarn reveals that it surpasses or closely matches human-curated conversations in fluency, relevance, and grammar. These findings challenge the notion that synthetically generated datasets are inherently inferior, positioning ConvKGYarn as a promising approach to producing high-quality conversational data in a repeatable and scalable manner.

5.3 Quantitative Analysis — Parametric Knowledge Evaluation

We evaluate LLMs on 100 examples from both *General* and *Related* sets, with ConvKGYarn generating conversations across all configurations. This consistent fact set enables confounder-free hypothesis testing, allowing analysis of LLM effectiveness with specific variables like typos in text interactions or combined deixis and disfluencies in voice interactions.

Figure 5 in Appendix A illustrates our iterative LLM evaluation process. We prepend each turn with the gold conversational history, using a prompt designed for accurate and relevant responses. The prompt allows Pythonic list-form answers for multiple valid responses and “NA” returns for low-confidence situations.

We tested GPT_{3.5} and GPT₄, using GPT₄ as a judge for binary rating of predictions (see Figure 6 in Appendix A). Our evaluation prompt systematically assesses response correctness, comparing candidate answers against gold standards for each turn. The scoring is 1 for correct responses, 0 otherwise, with list answers scored 1 if any candidate matches a gold answer.

This process respects conversation order, providing scores corresponding to turn sequences. It offers quantifiable metrics on LLM effectiveness, addressing limitations of F1 and EM scores to account for aliases and variations in LLM answers.

Table 3 presents GPT₄-EVAL results for *General* and *Related* settings, including mean scores at turn and conversation levels, and refusal rates (NA

Interaction	Setting			GPT _{3.5}			GPT ₄		
	Deixis	Disfluency	Typo	Mean (<i>Turn</i>)	Mean (<i>Conv.</i>)	NA Ratio	Mean (<i>Turn</i>)	Mean (<i>Conv.</i>)	NA Ratio
Voice	✗	✗	-	0.246 / 0.326	0.234 / 0.323	0.485 / 0.304	0.301 / 0.391	0.292 / 0.387	0.352 / 0.252
Voice	✗	✓	-	0.250 / 0.349	0.236 / 0.346	0.434 / 0.272	0.320 / 0.412	0.307 / 0.407	0.329 / 0.232
Voice	✓	✗	-	0.261 / 0.305	0.244 / 0.303	0.440 / 0.312	0.299 / 0.374	0.288 / 0.370	0.333 / 0.269
Voice	✓	✓	-	0.261 / 0.306	0.254 / 0.304	0.432 / 0.276	0.299 / 0.384	0.290 / 0.381	0.340 / 0.244
Text	✗	-	✗	0.246 / 0.333	0.233 / 0.329	0.459 / 0.276	0.316 / 0.371	0.294 / 0.366	0.335 / 0.285
Text	✗	-	✓	0.220 / 0.279	0.199 / 0.277	0.513 / 0.352	0.265 / 0.347	0.242 / 0.346	0.451 / 0.350
Text	✓	-	✗	0.239 / 0.307	0.221 / 0.302	0.445 / 0.306	0.269 / 0.361	0.248 / 0.355	0.385 / 0.309
Text	✓	-	✓	0.201 / 0.220	0.179 / 0.219	0.519 / 0.433	0.222 / 0.290	0.201 / 0.285	0.479 / 0.396

Table 3: The effectiveness based on the GPT₄-EVAL metric of two models GPT_{3.5} and GPT₄ when evaluated against variants of the *General* and *Related* sets (scores separated by /). Note that all settings for a particular set are grounded on the same facts.

Ratio). This comprehensive evaluation provides insights into LLM effectiveness across various conversational nuances and configurations.

Our analysis reveals several key findings. GPT₄ consistently outperforms GPT_{3.5}, likely due to its enhanced capabilities, more extensive training data, and refined instruction fine-tuning. The lower refusal rate for GPT₄ suggests more comprehensive information retention in its parameters.

The impact of voice versus textual interaction on effectiveness is inconclusive when not compounded by other linguistic phenomena. However, in the presence of deixis, there is a slight advantage for voice interactions, suggesting easier referent resolution in spoken queries with more contextual clues.

Surprisingly, disfluencies in voice interactions have a negligible or slightly beneficial effect, indicating LLMs’ growing ability to filter irrelevant signals and focus on core information needs. As expected, typos negatively impact both models’ effectiveness, highlighting their sensitivity to correct spelling for question comprehension and processing.

These results offer a nuanced understanding of LLM effectiveness in conversational factoid question-answering across diverse settings. We argue that such comprehensive evaluation across various configurations is crucial for developing a thorough assessment of system effectiveness, which ConvKGYarn enables at scale.

6 Conclusions

In this paper, we introduced ConvKGYarn, a novel framework to generate dynamic and scalable conversational datasets for KGQA. Our system leverages the structured representation of KGs to produce conversational datasets that can adapt to the evolving information needs of the user and knowl-

edge captured by KGs. Our extensive evaluations demonstrate that ConvKGYarn effectively generates well-configured high-quality KGQA datasets. By conducting rigorous qualitative and quantitative tests, we showcased that the datasets generated are versatile across various conversational scenarios, allowing us to test models on their effectiveness with different facets of user interactions and linguistic phenomena.

Furthermore, psychometric analyzes highlighted that the conversations generated from ConvKGYarn were comparable to those from traditional human-curated datasets, scoring highly on the metrics of relevance, fluency, cohesiveness, and grammar (when targeting these attributes) while being a few orders of magnitude larger in scale. An important finding from our work is the adaptability of ConvKGYarn to handle different types of interactions, such as text and voice, by appropriately configuring conversations to fit criteria and attributes such as deixis, disfluencies, and typos. In addition, our system’s ability to dynamically integrate updates from KGs ensures that the conversations remain current and factually accurate, addressing one of the significant challenges in existing conversational KGQA datasets.

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

Acknowledgement

We would like to thank Anil Pacaci, Simone Conia and Varun Embar for insightful discussions surrounding the choices 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*, pages 4661–4672, Online. 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*, pages 11922–11938, Abu Dhabi, United Arab Emirates. 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*, page 729–738, New York, NY, USA. 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*, pages 1612–1634, Singapore. 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*, 34(8):3549–3568.
- 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*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. 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.*, 54(4).
- Iliia 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*, pages 76–87, Tokyo, Japan. 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*, pages 119–126.
- Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docTTTTTquery.
- 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)*, pages 2463–2473, Hong Kong, China. 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 Sharifmoghammad, and Jimmy Lin. 2023a. RankVicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv:2309.15088*.
- Ronak Pradeep, Sahel Sharifmoghammad, 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*, 14(4):289–444.
- 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)*, pages 601–614, Online only. Association for Computational Linguistics.

- Neil Stewart, Gordon D A Brown, and Nick Chater. 2005. Absolute identification by relative judgment. *Psychol. Rev.*, 112(4):881–911.
- 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 Khoubisrat, 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 templated 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 7.

Figure 5 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 GPT₄ as a judge to rate the predictions in a binary fashion, as depicted in Figure 6.

B Human Annotation Process

In this section, we provide in-depth details on ConvKGYarn’s human annotation process used during the evaluation tasks.

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 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:

- Conversation A. The first conversation better meets the criteria for the given dimension compared to the second conversation.

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’), ...]

GPT4: [(‘P412’, ‘voice type’), ...]

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

- 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 × 1,000 conversations × 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 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?

GPT₄: *Ed Helms*

USER: Ummm, who was, hmm, its director?

GPT₄: *Alastair Fothergill*

Figure 5: Example Interaction for GPT_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 Ratings: <pythonic list of 0s/1s> where the list's 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
GPT₄: [1, 0, 1]
 ⋮

Figure 6: Prompt for GPT₄-eval of ConvKGYarn.

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

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.

- Familiarize** yourself with the grading metrics in Grading Information.
- Read** the conversation between Person 1 and Person 2.
- 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:

- Fluency
- Relevancy
- Response Diversity
- 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 location within
System 2: Pomona

Turn 2

System 1: the area of this place
System 2: 230 acre

Turn 3

System 1: the designation of heritage in this populated place
System 2: National Register of Historic Places listed place

Turn 4

System 1: the location of this populated place
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

Question 2
What is the **Relevancy** of the dialogue?
1 2 3 4 5

Question 3
What is the **Response Diversity** of the dialogue?
1 2 3 4 5

Question 4
What is the **Grammar** of the dialogue?
1 2 3 4 5

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.

- Familiarize** yourself with the grading metrics in Grading Information.
- Read** the conversation between Person 1 and Person 2.
- 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:











- Fluency
- Relevancy
- Response Diversity
- 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 Dialogue 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	<p>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.</p> <p>Text is basic, often fragmented, and may miss key connecting words.</p>	<p>Translated Question: "Biggest mountain what?"</p> <p>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.</p> <p>In addition, "biggest" would not be commonly be used to ask about the tallest mountain.</p>
2 - Elementary	<p>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.</p> <p>The text is more structured than the beginner level but might still lack proper phrasing.</p>	<p>Translated Question: "What mountain biggest?"</p> <p>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.</p>
3 - Limited	<p>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.</p> <p>The text is clearer, conveying straightforward information with better structure. Word choices are more natural as well.</p>	<p>Translated Question: "What is the mountain with the maximum elevation on Earth?"</p> <p>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.</p> <p>However, it sounds artificial, using technically correct language, that wouldn't be commonly used. More common variants are "tallest" or the "highest".</p>
4 - Professional	<p>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.</p> <p>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.</p>	<p>Translated Question: "Which is the tallest mountain in the world?"</p> <p>Reason: This would be a perfectly fine way to phrase the source question. However, there is only one discrepancy, that differs from truly natural and local translations. Instead of "which", most people would use "what".</p>
5 - Native	<p>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.</p> <p>Text is straightforward, fully natural, and effortlessly conveys the intended information or question. Words choices are native as well.</p>	<p>Translated Question: "What is the tallest mountain in the world?"</p> <p>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.</p>

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.	<p>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?"</p> <p>Reason: None of these questions correlate with each other on the theme or information.</p>
2 - Slightly Relevant	Some attempts at connection between topics, but many turns in the conversation feel forced or out of place.	<p>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?"</p> <p>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.</p>
3 - Moderately Relevant	Most turns in the conversation relate to a central topic, but there are occasional drifts into unrelated subjects.	<p>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?"</p> <p>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.</p>
4 - Highly Relevant	Nearly all turns in the conversation have clear ties to a main topic or theme, with minimal deviation.	<p>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?"</p> <p>Reason: Technically, each turn in the conversation has the connection to the next. However, the connections do not seem too natural in a conversation.</p>
5 - Completely Relevant	Every turn in the conversation seamlessly flows from one to the next, maintaining a single, clear focus throughout.	<p>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?"</p> <p>Reason: Each turn in the conversation relate to each other, and the entire conversation has a central theme and intuitive flow.</p>

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.	<p>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?"</p> <p>Reason: Only asking surface level questions about Mona Lisa.</p>
2 - Below Average Diversity	Shows slight variation in entities or topics, but the types of questions remain largely consistent or predictable.	<p>Example: "What is the Mona Lisa? Who painted the Mona Lisa? Who painted The Last Supper? When was The Starry Night painted?"</p> <p>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.</p>
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.	<p>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?</p> <p>Reason: Has more diversity of entities and the type of questions that are asked across the different entities themselves. But it is stuck in the smaller realm of art.</p>
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.	<p>Example: "What is the Mona Lisa? Leonardo Da Vinci's famous artworks? Other influential art figures in the Renaissance era?"</p> <p>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.</p>
5 - High Diversity	Demonstrates a wide spectrum of question types related to various entities, with seamless transitions and consistent coherence throughout the conversation.	<p>Example: "What is the Mona Lisa? Famous Renaissance painters in Europe? Is Beethoven Renaissance music? Can you name some contemporary artists inspired by classical art?"</p> <p>Reason: It traverses various entities, and asks unique questions about each of them, while still in the bounds of logical flow.</p>

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.	<p>Translated Question: "eiffel tower were is?"</p> <p>Reason: The overall structure of the sentence is incorrect. In addition, Eiffel Tower was not capitalized. "Where" is not correctly spelled.</p> <p>Due to the errors, the question might not be understandable.</p>
2 - Novice	Has multiple grammatical mistakes but the central question or point is discernible. Has some spelling errors.	<p>Translated Question: "Where Eiffel Tower located."</p> <p>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.</p> <p>You can understand the question, but it is obvious there are mistakes.</p>
3 - Intermediate	Displays occasional grammatical errors but the message remains clear. Has only a couple spelling errors.	<p>Translated Question: "Where are the Eifel Tower location?"</p> <p>Reason: Eiffel Tower is incorrectly spelled, and "where are" should be "where is" due to it being singular.</p> <p>You can easily understand the question; however, there are a couple minor errors.</p>
4 - Advanced	Demonstrates very few and minor grammatical errors that don't hinder comprehension. Potentially has a single spelling error.	<p>Translated Question: "Where does the Eiffel Tower located?"</p> <p>Reason: The sentence structure is correct, but there is a minor mistake of using "does" instead of "is".</p>
5 - Expert	Showcases exemplary grammar without errors.	<p>Translated Question: "Where is the Eiffel Tower located?"</p> <p>Reason: The translated question has correct grammar, consisting of correct sentence structure, punctuation and capitalization</p>

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.