

Automatic Generation of Inference Making Questions for Reading Comprehension Assessments

Wanjing Anya Ma
Stanford University *
Stanford, CA, USA
wanjingm@stanford.edu

Michael Flor
ETS Research Institute
Princeton, NJ, USA
MFlor@ets.org

Zuwei Wang
ETS Research Institute
Princeton, NJ, USA
zwang@ets.org

Abstract

Inference making is an essential but complex skill in reading comprehension (RC). Some inferences require resolving references across sentences, and some rely on using prior knowledge to fill in the detail that is not explicitly written in the text. Diagnostic RC questions can help educators provide more effective and targeted reading instruction and interventions for school-age students. We introduce a taxonomy of inference types for RC and use it to analyze the distribution of items within a diagnostic RC item bank. Next, we present experiments using GPT-4o to generate bridging-inference RC items for given reading passages via few-shot prompting, comparing conditions with and without chain-of-thought prompts. Generated items were evaluated on three aspects: overall item quality, appropriate inference type, and LLM reasoning, achieving high inter-rater agreements above 0.90. Our results show that GPT-4o produced 93.8% good-quality questions suitable for operational use in grade 3-12 contexts; however, only 42.6% of the generated questions accurately matched the targeted inference type. We conclude that combining automatic item generation with human judgment offers a promising path toward scalable, high-quality diagnostic RC assessments.

1 Introduction

Inference-making is an essential yet cognitively demanding skill in reading comprehension (RC) (O'Brien et al., 2015; Kintsch, 1998). Inferences are necessary for establishing both local and global coherence within the mental representation of a text (Graesser et al., 1994). Local inferences connect information across sentences using cohesive devices such as anaphors or category exemplars—for example, in "Bette gulped down the drink. The cold water was very refreshing," the reader infers that *the drink* refers to *cold water* (Cain, 2022, p. 307).

Global inferences, on the other hand, rely on the reader's prior knowledge to fill in missing details required to make sense of the text—for example, in "The campfire started to burn uncontrollably. Tom grabbed a bucket of water" (Bowyer-Crane and Snowling, 2005, p. 192), the reader infers that Tom intended to put out the fire, based on the knowledge that water extinguishes fire. While skilled readers often generate inferences automatically as they engage with text (Thurlow and van den Broek, 1997), children who struggle with comprehension frequently have difficulty constructing these inferences (Cain et al., 2001).

Providing diagnostic information about specific types of inference-making deficits that hinder comprehension can empower educators to provide more effective and targeted reading instruction and intervention (Bowyer-Crane and Snowling, 2005; Bayat and Çetinkaya, 2020). To achieve this, we need RC assessments that specifically target inference-making types. At the same time, we want to develop scalable item generation methods to enable multi-time testing, monitoring reading development over time. Previous work has demonstrated the ability of large language models (LLMs) to generate effective RC questions (Uto et al., 2023; Säuberli and Clematide, 2024). However, whether LLMs can reliably produce questions that target specific inference types remains unclear.

Our research is grounded in a real-world diagnostic assessment of reading skills for students in grades 3 through 12 (Sabatini et al., 2019). The assessment was originally developed at ETS and recently commercialized as ReadBasix. It leverages the science of reading to assess foundational reading skills, such as word recognition and decoding, as well as more complex ones such as RC. In the RC subtest, a student will usually read 4 expository passages and answer multiple-choice questions associated with the passages. The subtest takes about 30 minutes to complete. Like any

*Work done while at ETS Research Institute

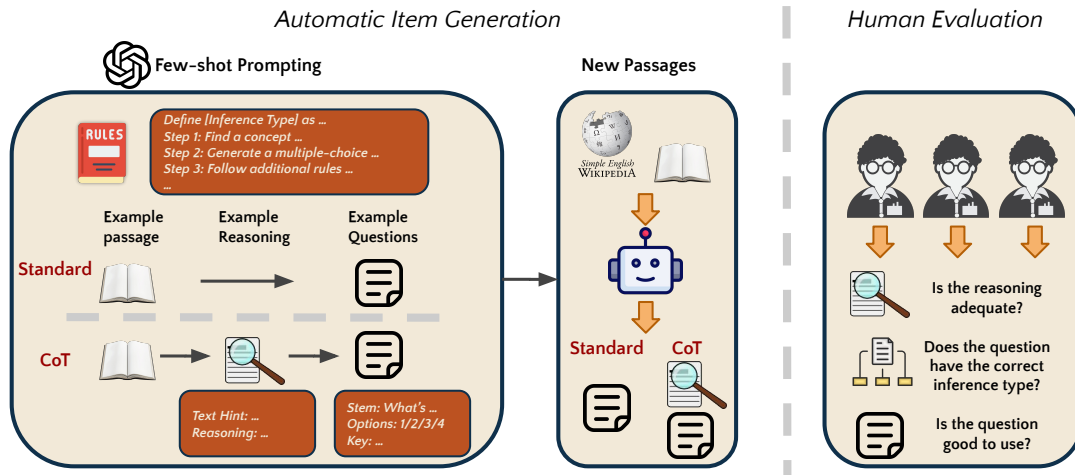


Figure 1: **Overview of automatic item generation and human evaluation.** We use GPT-4o to generate bridging-inference RC items for given reading passages via few-shot prompting, comparing conditions with and without chain-of-thought prompts. We prompt each inference type separately: pronominal bridging, text-connecting, and gap-filling inferences. Human evaluation focuses on general item quality, inference type appropriateness, and LLM rationales.

large-scale reading assessment, there is an ongoing need for more items. To address this demand, we aim to leverage automatic item generation to create new items based on curated passages, and evaluate the quality of these items before collecting student performance data to make them operational.

For the purpose of automatic item generation, as illustrated in Figure 1, we first conducted a literature review on inference-making in the reading comprehension and natural language processing (NLP) text comprehension literature. We developed a taxonomy of inference-making questions, with a focus on bridging inference. We validated this taxonomy by annotating an operational item bank of expert-written RC questions, confirming bridging inference as an important and widely covered sub-construct. Next, we curated six expository passages and manually wrote multiple-choice RC questions for each inference type based on our taxonomy. These examples were then used to prompt GPT-4o (Hurst et al., 2024) via few-shot prompting to generate bridging-inference questions for new reading passages, comparing conditions with and without chain-of-thought (CoT) prompting (Wei et al., 2022). Finally, three human experts evaluated the quality of the generated questions along three dimensions: overall item quality ¹, appropriate inference type, and whether GPT-4o provided

¹The evaluation of overall item quality does not include whether an item is of the required inference type, which is an extra-evaluation. See Table 2 for more details.

satisfactory reasoning for generating the question. Our results show that LLMs can produce 93.8% good-quality questions suitable for operational use in grade 3-12 contexts; however, only 42.6% of the generated questions accurately match the targeted inference type. Nevertheless, the overall coverage of inference types closely mirrors what we observe in our operational item bank. We conclude that combining automatic item generation with human judgment offers a promising path toward scalable, high-quality diagnostic RC assessments.

In summary, we make the following contributions in this paper:

1. We develop and validate a taxonomy for inference-making questions used in multiple-choice RC assessments, and demonstrate its value for future item development.
2. We introduce a novel NLP task where language models generate RC questions targeting specific inference types, providing a new way to assess their reasoning abilities. The training item bank will be released for replication and benchmarking.
3. We demonstrate GPT-4o’s potential in generating RC questions for operational use and its limitations in accurately generating specific types of inference questions.

2 Related Work

2.1 Question generation for reading comprehension assessments

Automatic question generation is a well-established task in NLP, especially within educational applications, to reduce the high costs of manual question authoring and to ensure a steady supply of new, high-quality items (Kurdi et al., 2020). Early approaches rely on rule-based or template-based methods (Araki et al., 2016; Flor and Riordan, 2018), as well as the use of discourse connectives to generate questions (Agarwal et al., 2011). Later approaches extensively used neural systems for question generation (Mulla and Gharpure, 2023). More recent work demonstrates that LLMs hold promise in generating high-quality RC questions, using techniques such as fine-tuning (Uto et al., 2023; Perkoff et al., 2023; Ghanem et al., 2022; Ashok Kumar et al., 2023; Rathod et al., 2022; Stasaski et al., 2021), zero-shot or few-shot prompting (Säuberli and Clematide, 2024; Attali et al., 2022), and Chain-of-Thought prompting (Kulshreshtha and Rumshisky, 2022). Some of these studies have also explored the generation of more complex, "deeper" questions—those that target underlying reasoning processes (Ghanem et al., 2022; Poon et al., 2024) or hinge on specific inference steps for accurate responses (Araki et al., 2016). Within the domain of automated Question Answering, the notion of *multi-hop questions* has gained attention, as questions relating different parts of a document require multi-step reasoning (Mavi et al., 2024).

We note that prior studies have largely treated reading comprehension as a single, undifferentiated construct even though comprehension requires different types of inferences. Recent work has begun to develop taxonomies of RC and annotate question types to enable more controllable generation (Xu et al., 2022; Li and Zhang, 2024; Hwang et al., 2024). However, to our knowledge, no existing work has systematically addressed question generation based on specific types of inference. We believe that the capability to generate different types of inference questions will provide more diagnostic insights for educators. Our work is a first step toward filling this gap.

2.2 Bridging inference as an NLP task

The NLP community has long tackled text comprehension challenges, including bridging infer-

ence. Prior work has focused on corpus-based bridging anaphora recognition and resolution using annotated resources such as ISNotes and BASHI (Rösiger, 2018; Hou et al., 2018; Hou, 2020). Neural models have been developed to jointly learn mention representations and bridging relations (Pandit and Hou, 2021; Kobayashi et al., 2022). In the recently developed IdentifyMe benchmark for resolving nominal and pronominal mentions across long contexts (Manikantan et al., 2024), GPT-4o outperforms other LLMs, achieving 81.9% accuracy and demonstrating strong referential capabilities. With the rise of LLMs, research increasingly shifts toward evaluating LLMs' general reasoning capabilities (Brown et al., 2020; Wei et al., 2022). In our education application, we investigate whether LLMs truly possess the reasoning ability required for bridging inference, particularly through the lens of a question generation task.

3 Taxonomy of Inference Questions

3.1 Development of Taxonomy

Inferences can be categorized into bridging inferences, elaborative inferences, predictive inference, emotional inference, etc (Graesser et al., 1994; Schmalhofer et al., 2002; Singer and Remillard, 2004; van den Broek et al., 2015). To manage the scope of our interest, we focus on bridging inference which connects information in a text. Bridging inferences contribute to text coherence by allowing the reader to identify the connections among concepts and ideas in the text (Singer et al., 1992; Singer and Remillard, 2004) or bridges (Haviland and Clark, 1974) among the propositions underlying the discourse. A bridging inference is needed when the reader cannot retrieve a referent for the given information of the current sentence from either working memory or long-term memory.

Table 1 shows the taxonomy of inference making questions for diagnostic RC assessments, along with the examples. The first type is **pronominal**, and it has two variants. Simple pronominal asks for a direct pronoun resolution, such as "In the sentence, whom does 'he' refer to?" This is different from the second subtype: **pronominal bridging**, which requires the reader to use the pronoun as a hint to bridge sentences and answer the question. The third type **text-connecting** requires test takers to connect two explicitly stated components in a text, and usually the bridge are noun phrases. The last type is **gap-filling**, which requires readers to

Types	Definitions	Examples
Pronominal	Direct pronoun resolution.	Like "To whom 'he' refers?", "What does 'this' represent?"
Pronominal Bridging	Use pronoun as a hint to bridge sentences.	Text snippet: <i>Ships have carried passengers since prehistoric times. That is the first kind of public transportation.</i> Question: <i>What was the first kind of public transportation in history?</i> Answer: <i>ships</i> Reasoning: <i>The pronoun "That" refers to "ships" in the previous sentence.</i>
Text-Connecting	Connecting two explicitly stated components in a text, typically through a noun phrase.	Text snippet: <i>Public transportation is good for the environment. When many people use the same vehicle, fewer cars are on the road. Fewer cars make less pollution.</i> Question: <i>Why is public transportation good for the environment?</i> Answer: <i>Because it causes less pollution</i> Reasoning: <i>"Fewer cars" links to "public transportation" from the previous sentence in a causal relationship.</i>
Gap-Filling	"Incorporating information outside of the text, i.e., general knowledge, with information in the text to fill in missing details." (Cain and Oakhill, 1999, p.490)	Text snippet: <i>White pizza uses no tomato sauce, often substituting pesto or dairy products such as sour cream. Most commonly, its toppings consist only of mozzarella and ricotta cheese drizzled with olive oil and basil and garlic.</i> Question: <i>What is a possible reason "White pizza" gets its name?</i> Answer: <i>It doesn't have tomato sauce</i> Reasoning: <i>Readers need to use common sense to fill in the gap that "no tomato sauce" means the color of the pizza is not red.</i>

Table 1: Taxonomy of inferences for Reading Comprehension questions.

incorporate information from outside of the text with information in the text to fill in some missing details. More examples based on the taxonomy are included in Appendix A.

3.2 Validation of Taxonomy

With the newly developed taxonomy, we annotated the RC items in an in-house item-bank. The item-bank has 192 expert-written multiple-choice RC questions for 24 expository reading passages. These passages vary in difficulty from Grade 3 to Grade 12. Our primary focus was to classify the types of bridging inferences, but we also annotated questions that are not in our main scope of interest. For example, there are some **factual/literal** questions, for which a test taker can directly find information from the text without involving inference; **vocabulary** questions that directly assess the vocabulary knowledge, and other comprehension questions that do not require bridging inferences.

Two of the co-authors classified items independently, following the annotation guideline (see Appendix B). The two coders provided the same coding of the type of inference on 86% of the items, with kappa = 0.83, indicating high agree-

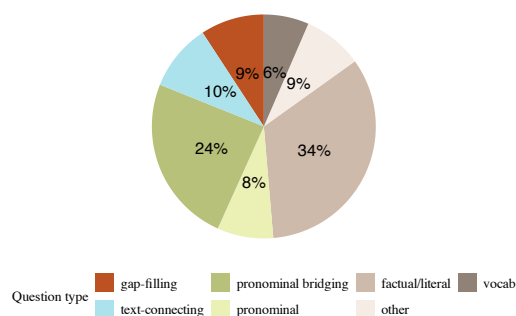


Figure 2: Distribution of different inference types in an operational reading comprehension item bank.

ment. Based on our annotation results shown in Figure 2, we find that bridging inference questions account for 51% of the RC items in the item bank, suggesting bridging inference is an important sub-construct in this RC assessment. Among the bridging inference questions, pronominal bridging (24%) is the most dominant type, followed by text-connecting (10%), gap-filling (9%), and pronominal questions (8%). The high level of agreement supports the validity of the newly developed taxonomy, which we see as an important contribution—

providing a road-map for both item development and future research.

4 Automatic Item Generation and Human Evaluation

Figure 1 presents the overview of our automatic item generation pipeline.

4.1 Training Questions

Due to test security considerations we can not use texts and items from our operational item bank as examples to prompt LLMs. Thus, we created our example item bank which is publicly available for replication efforts². We adapted 6 new expository passages from Simple English Wikipedia³ (passage length ranges from 342 to 508 words, average 438) and for each passage we manually created 2-4 items for each type of inference. Each question contains a **stem**, four **options**, and an answer **key** indicating which option is correct. We also included our thought process in the item generation: **text hint** includes the relevant text from the passage where required inference will be made, and **reasoning** is a short explanation why this question belongs to the requested type. In total, we wrote 19 pronominal bridging, 23 gap-filling, and 16 text-connecting questions.

4.2 Few-shot Prompting

We used the GPT-4o model (2024-04-01-preview) to generate multiple-choice RC questions based on passages we supplied to the model. To prioritize accuracy and reproducibility in item generation, we set the temperature parameter to 0. We explored the frequency penalty parameter from 0 to 0.3, with 0.2 proving optimal as it could consistently generate three diverse RC items without compromising their quality.

Few-shot prompting techniques were used and the prompts were iteratively refined over six rounds. Most adjustments focused on improving the concreteness of the question-writing steps to better guide the model. In this paper, we only report the final iteration of item generation in which we experimented with four different prompting conditions: standard prompting with 4 (or 6) passages and examples, and chain-of-thought prompting with 4 (or 6) passages and examples with text hint and reasoning. With this set-up, we investigated whether

²<https://github.com/maafiah/InferenceQuestionsAQG>

³<https://simple.wikiedia.org>

Task: Given a passage, you are going to generate pronominal bridging inference questions.

Follow these steps to answer the user queries.

Step 1 - find a pronoun (it, they, she, he, which, that, etc) in the passage that is connecting AT LEAST 2 or 3 sentences. The pronoun should be crucial to bridge meaningful information from the passage such as a fact, a cause, a result, or a feature.

Step 2 - based on the pronoun and its reference, generate a multiple-choice question with three distractors. The question should use the pronoun and its reference as hints to connect information between sentences.

Step 3 - follow additional rules when writing the questions: 1) do not ask a question that requires background knowledge to answer. 2) do not ask a question that directly asking "what does XX refer to". 3) lightly paraphrase the question and option without introducing new inference. 4) do not write correct answer longer than the distractors.

Step 4 - iterate this process for 2 times to get 3 different questions.

Step 5 - Output by following the exact format as examples so that it can be directly converted to csv format (do not have any title like (**questions**)). Include all the sentences required in the 'Text Hint' and output your thought process in the 'Reasoning'.

Here are some example passages and example questions:

*****Given passage:*****
A greenhouse is a building where plants such as flowers and vegetables are grown. It usually has a glass ...

*****Examples:*****
 PassageName\Inference Type
 \Text Hint\Reasoning
 \Stem\Option 1\Option 2\Option 3\Option 4\Key
 Greenhouse\pronominal bridging
 \A greenhouse is a building where plants such as flowers and vegetables are grown. It usually has a glass or translucent plastic roof.\the pronoun "it" refers to "greenhouse" in the previous sentence.
 \According to the passage, what can have translucent plastic roofs?\backyards\living spaces\greenhouses \botanic gardens\3
 ...

*****New Passage:*****
Parallax is the perceived change in position of an object seen from two different places ...

Figure 3: **Few-shot prompt for generating pronominal bridging inference questions.** The system prompt (beige background) defines the inference type and outlines expert-inspired steps. Training examples (provided in the prompt) follow. In the standard condition, only the question and answer key (green) are shown; in the CoT condition, text hints and reasoning (blue) are also included. A new passage is provided in the user prompt (orange background) to generate new questions.

increasing the training examples or using the CoT strategy would improve the quality of generation. Moreover, we further evaluated if the output rea-

Criterion	Annotation Guidelines
General item quality	1: If the generated item satisfies all of the following: (a) The correct answer is fully correct; (b) Distractors are not confusing and are clearly incorrect; (c) The question is developmentally appropriate and safe for Grades 3–12. 0: If any requirement is not met. Provide an explanation in the "Note" field.
Inference-type accuracy	1: If the generated item matches the requested inference type. 0: If not. Output inference type, one of: gap-filling / pronominal bridging / text-connecting / factual or literal.
Reasoning quality	1: If the generated thought process fulfills both of the following: (a) The "Reasoning" is adequate and relevant to the requested inference type; (b) The "Text Hint" includes all the sentences required to answer the item correctly. 0: If either condition is not satisfied.

Table 2: Annotation guidelines for evaluating the generated items.

soning process was adequate for this specific task.

Figure 3 shows an example prompt for generating reading comprehension questions targeting pronominal bridging inference (see Appendix C for more details). In the system prompt, we first instructed GPT-4o to identify pronominal bridging relationships, then directed it to generate a multiple-choice question, guided by additional rules to ensure item quality. We included several training examples in the prompt—either 4 or 6 passages with corresponding questions, depending on the generation condition. For the Standard condition, no text hints or reasoning were provided in the training examples. In the CoT condition, both text hints and reasoning were provided, prompting the model to generate them in the output. In the user prompt, we provided a new passage for GTP-4o to generate items from.

We curated a total of 10 new passages adapted from Simple Wikipedia, which were comparable in length and format to the example passages. For each passage and inference type (pronominal bridging, text-connecting, and gap-filling), we independently applied the prompting procedure, instructing GPT-4o to generate three unique questions per combination. For text-connecting and gap-filling—where question construction can be more challenging—we included an additional rule: "Do not force additional questions if no suitable locations can be found." Across the four prompting conditions, we generated a total of 357 questions, 180 of which were produced under the CoT condition and therefore included text hints and reasoning in the output.

4.3 Human Evaluation

To evaluate the quality of the generated RC items, we developed an evaluation rubric (see Table 2). Three authors used items from prior iterations of the generation process and complete several practice rounds and discussion before finalizing the rubric. The rubric is designed to directly address our core research questions:

RQ1: Can LLMs generate high-quality RC items with appropriate distractors suitable for inclusion in an operational item bank?

RQ2: Do the generated RC items align with the requested bridging inference type?

RQ3: How well can LLMs reason about their generation process?

In the evaluation phase, the three authors, who are experts in reading assessment questions, independently annotated all 357 generated items. The agreement was high for general item quality (RQ1), with percent agreement ranging from 87–90%. However, reaching consensus on the inference type (RQ2; 69–70%) and reasoning quality (RQ3; 65–71%) proved more challenging—consistent with prior findings that reasoning-related judgments are inherently difficult to rate (Stasaski et al., 2021).

To address this, we conducted a second round of annotation. In this phase, each rater independently reviewed only the items where their initial rating differed from the other two and decided whether to adjust the rater’s original score. Following this adjustment, inter-rater agreement improved substantially. The final results of percentage agreement and Fleiss’ kappa are shown in Table 3. Our

<p>Requested Type: Gap-filling</p> <p>Text Hint: The main way carbon gets taken out of the atmosphere is by photosynthesis by living organisms.</p> <p>Reasoning: requires common sense to know that photosynthesis is performed by plants.</p> <p>Question: Which organisms play a crucial role in removing CO2 from the atmosphere?</p> <p>Options: animals/ bacteria/ plants/ fossil fuels</p> <p>Key: 3</p> <p>Rating: high quality, correct inference type, and correct reasoning.</p>	<p>Requested Type: Text-Connecting</p> <p>Text Hint: Doughnuts are often eaten in the morning, along with a cup of hot coffee. They are sold at doughnut shops, bakeries, or grocery stores.</p> <p>Reasoning: "doughnuts" and "doughnut shops" are linked thematically."</p> <p>Question: Where can people buy doughnuts?</p> <p>Options: At a coffee shop/ At a doughnut shop/ At a restaurant/ At a candy store</p> <p>Key: 2</p> <p>Rating: low quality because multiple keys can be correct. The question has incorrect inference and reasoning. The question can be categorized as pronominal bridging, as "they" refers to "doughnuts," or as factual/literal, since "doughnut shops" directly refers to places where doughnuts are sold.</p>
---	--

Figure 4: Examples of LLM-generated RC items via Chain-of-Thought prompting with 6 training passages. Left: high-quality; right: low-quality. Each output includes a text hint, a rationale, a multiple-choice question with four options, and an answer key. Human annotations are shown against a beige background.

Criterion	Agreement (%)	Fleiss' κ
General item quality	90–97	0.57
Inference-type accuracy	85–94	0.77
Reasoning quality	90–95	0.83

Table 3: Inter-rater agreement and Fleiss' κ for each evaluation criterion. Agreement is reported as a range based on three pairwise comparisons by three graders.

evaluation in the Results section were based on the majority votes for each item. For example, an item was treated as acceptable when at least two of the three raters rated it as good quality.

5 Results

Based on the proportion of accepted items by generation method (Table 4), we observe improved generation performance when increasing the number of training examples from four to six example passages in the prompt. However, our experiment does not show any clear advantage of Chain-of-Thought prompting over standard few-shot prompting. Furthermore, our results indicate no statistically significant differences in generation performance across the various prompting conditions. We summarize our key findings below.

LLMs can produce high-quality questions suitable for operational use. Based on the evaluation of general item quality, 87 out of 90 questions (96.7% in the CoT_6 condition) had good quality and were suitable for operational use in the Grade 3-12 educational context. The performance is com-

Generation Method	Num Items	General Item Quality	Inference Accuracy	Reasoning Quality
standard_4	88	0.932	0.409	
standard_6	89	0.955	0.461	
CoT_4	90	0.900	0.411	0.356
CoT_6	90	0.967	0.422	0.389
Total	357	0.938	0.426	0.372

Table 4: Proportion of accepted items by generation method—standard vs. chain-of-thought prompting (with text hints and reasoning), using 4 or 6 passages (12–18 examples). Highest scores per criterion are bolded; criteria are defined in Table 2.

parable to, if not better than, those reported in prior research evaluating overall item quality for RC assessments, which ranged from 75% to 90% (Kulshreshtha and Rumshisky, 2022; Uto et al., 2023; Säuberli and Clematide, 2024). Because of the differences between these studies, for a more informative comparison, we encourage future research to replicate our findings under similar conditions. Figure 4 presents one high-quality example and one low-quality example of the generated questions. We find that problems of unacceptable questions included multiple keys, introduction of new vocabulary, confusing wording of the question, etc.

Generating RC questions by specific inference type is a challenging NLP task. Although LLMs can generate high-quality RC items, their ability to produce questions targeting specific inference types remains limited. In the generation method yield-

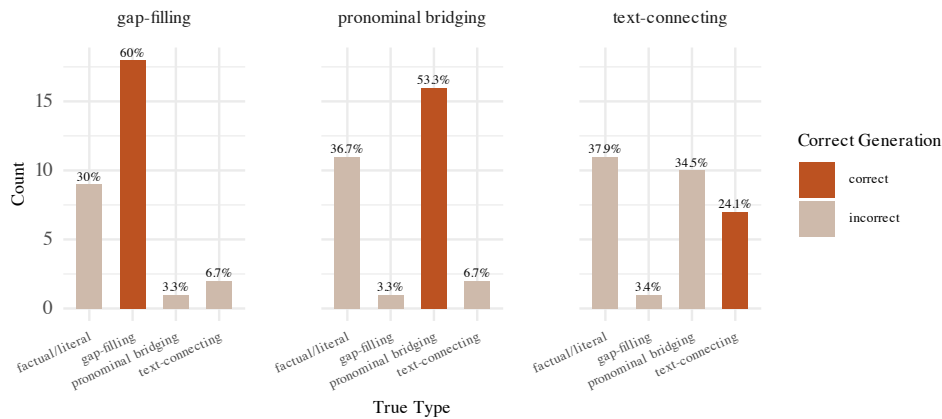


Figure 5: Human evaluation of inference-type accuracy. Each panel displays the distribution of true inference types corresponding to each requested inference type. The generation questions are obtained from the standard few-shot prompting with 6 training passages.

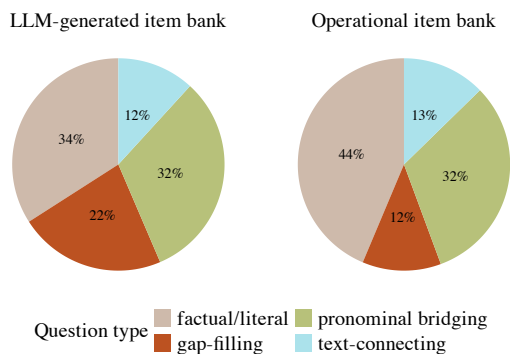


Figure 6: Comparison of item inference type coverage between the operational item bank and the LLM-generated item bank.

ing the best performance (standard_6), only 46.1% of the generated questions matched the requested inference type. As shown in Figure 5, gap-filling questions were the easiest to generate (60% match), followed by pronominal-bridging questions (53.3% match). In contrast, generating text-connecting questions proved particularly difficult, with an accuracy of only 24.1%. This pattern of generation difficulty aligns with the challenges faced by human experts (co-authors) when writing the training examples. We also find that 34.8% of the generated questions were factual or literal, requiring little inference. Moreover, GPT-4o provided adequate reasoning for only 38.9% of the items. This finding may explain the lack of performance gains when moving from standard prompting to CoT prompting. While prior work has shown that adding structured rationales can improve the accuracy of multi-

hop question generation (Säuberli and Clematide, 2024), we believe our task poses a more challenging test of an LLM’s reasoning ability.

Automatic item generation with human evaluation ensures the quality of diagnostic RC items.

From an application standpoint, we also examined how closely the distribution of inference types in the generated items resembled that of human-written items from our operational RC item bank. Interestingly, our analysis, shown in Figure 6, reveals that the overall distribution of inference types in the LLM-generated items closely matches that of our operational RC item bank. This means whereas GPT-4o failed to consistently produce individual items targeting specific inference types, the collect of items it generated somehow resembles the distribution of item types in our existing item pool. With some expert review, most of these items are suitable to use. Understanding the strengths and limitations of current LLM performance is important, particularly if we aim to rely on human evaluation to ensure quality and safety. The generation process is considerably more scalable than relying on human experts to write items manually. Despite current limitations, LLM-based item generation with our newly developed taxonomy offers a promising approach for educational applications.

6 Conclusion

This paper demonstrates our effort in leveraging a large language model to generate inference-making questions for a reading comprehension assessment. We developed a taxonomy of bridging inference questions based on existing literature and validated

it with empirical data from an operational test. The taxonomy focuses on three types of inferences: pronominal bridging, text connecting, and gap-filling. The taxonomy guided our manual creation of example comprehension questions, which were then used as training materials for GPT-4o to generate new items for the new passages. Our evaluation indicates that although GPT-4o can produce acceptable RC questions, its ability to generate questions aligned with specific inference types was limited. This limitation might stem from its limited capability in providing valid reasoning for the types of inferences. These results highlight the critical role of human evaluation when using LLMs for RC question creation. We propose that combining automatic item generation with human judgment offers a promising path toward scalable, high-quality diagnostic RC assessments.

Limitations

We provide preliminary evidence for the potential of GPT-4o in creating inference making reading comprehension questions. The following limitations should be addressed by future research.

We have a limited evaluation set. Our evaluation relies on 10 expository passages (based on Simple Wikipedia), restricting the generalizability of our findings to broader reading contexts or varied educational materials. Future research should incorporate more passages and of different genres, such as narratives.

We exclusively use GPT-4o. This study employed only one LLM, GPT-4o, which may limit insights into the potential effectiveness of other advanced reasoning models. Given the challenge of this reasoning task, future research should explore additional models. Because more advanced models may incur significantly higher costs, future research should also consider the balance between performance and affordability for an educational application.

Unclear effectiveness of Chain-of-Thought prompting. Our results show that generation quality improves with more example questions. However, our experiment does not show benefits from CoT prompting. This unexpected finding may result from our limited number of training examples. Future studies should expand the training data and possibly utilize large datasets, such as SQuAD (Rajpurkar et al., 2016) and FairytaleQA (Xu et al.,

2022). Future work should also explore more effective methods for integrating human-experts' rationales into the question generation process and explore how it affects the reasoning performance of LLMs (Zelikman et al., 2022).

General item quality is a broad metric. Our main goal is to generate RC items that target specific inference types, so we grouped other aspects like answer correctness and distractor plausibility under a broad "General Item Quality" metric. Still, there are important dimensions we didn't separate out—like item difficulty and whether it's appropriate for the target population. More specific metrics could help pinpoint where generation errors happen and how inference type and item difficulty might interact.

Future work should focus on item evaluation in real-world deployment. Our study did not include pilot testing in real-world settings to evaluate how the generated items perform with actual student responses. Student response data would allow for further examination of item bias, difficulty, and discrimination—critical steps before using the items for student scoring and making valid inferences about their abilities (Yeatman et al., 2024). Using LLM-simulated student responses to evaluate generated items is also an exciting direction that could help reduce—but not replace—the need for traditional item calibration (Zelikman et al., 2023; Lu and Wang, 2024; Liu et al., 2025).

Ethics Statement

Our study goal is to leverage LLMs to develop scalable and effective RC assessments to align with educational practice. We introduce a novel and meaningful NLP task: generating RC questions by inference type. While LLMs show promise for item development, we emphasize the importance of maintaining test security by avoiding training models on operational test items, and by ensuring the safety of content such as developmental appropriateness and the absence of problematic materials. In addition to existing automatic benchmarks, human evaluation by educational experts remains essential for item quality. Though beyond our current scope, we also highlight the need for ongoing monitoring of the generated items to detect scoring biases and ensure fairness in operational use.

Acknowledgments

We appreciate the reviewers for their helpful feedback on the manuscript. This study was made possible by the following research grant awarded by the Institute of Education Sciences, U.S. Department of Education, through R305F100005. Opinions, findings, and conclusions in this paper do not necessarily reflect the views of IES or ETS.

References

- Manish Agarwal, Rakshit Shah, and Prashanth Manem. 2011. Automatic question generation using discourse cues. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.
- Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1125–1136.
- Nischal Ashok Kumar, Nigel Fernandez, Zichao Wang, and Andrew Lan. 2023. [Improving reading comprehension question generation with data augmentation and overgenerate-and-rank](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 247–259, Toronto, Canada. Association for Computational Linguistics.
- Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903077.
- Nihat Bayat and Gökhan Çetinkaya. 2020. The relationship between inference skills and reading comprehension. *Education and Science*.
- Claudine Bowyer-Crane and Margaret J Snowling. 2005. Assessing children’s inference generation: What do tests of reading comprehension measure? *British journal of educational psychology*, 75(2):189–201.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Kate Cain. 2022. Children’s reading comprehension difficulties. *The science of reading: A handbook*, pages 298–322.
- Kate Cain and Jane V Oakhill. 1999. Inference making ability and its relation to comprehension failure in young children. *Reading and writing*, 11:489–503.
- Kate Cain, Jane V Oakhill, Marcia A Barnes, and Peter E Bryant. 2001. Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & cognition*, 29(6):850–859.
- Michael Flor and Brian Riordan. 2018. A semantic role-based approach to open-domain automatic question generation. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 254–263.
- Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer von der Ohe, and Alona Fyshe. 2022. [Question generation for reading comprehension assessment by modeling how and what to ask](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2131–2146, Dublin, Ireland. Association for Computational Linguistics.
- Arthur C Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371.
- Susan E Haviland and Herbert H Clark. 1974. What’s new? acquiring new information as a process in comprehension. *Journal of verbal learning and verbal behavior*, 13(5):512–521.
- Yufang Hou. 2020. [Bridging anaphora resolution as question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online. Association for Computational Linguistics.
- Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Kevin Hwang, Kenneth Wang, Maryam Alomair, Fow-Sen Choa, and Lujie Karen Chen. 2024. Towards automated multiple choice question generation and evaluation: aligning with bloom’s taxonomy. In *International Conference on Artificial Intelligence in Education*, pages 389–396. Springer.
- Walter Kintsch. 1998. *Comprehension: A paradigm for cognition*. Cambridge university press.
- Hideo Kobayashi, Yufang Hou, and Vincent Ng. 2022. [End-to-end neural bridging resolution](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 766–778, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Saurabh Kulshreshtha and Anna Rumshisky. 2022. Reasoning circuits: Few-shot multihop question generation with structured rationales. *arXiv preprint arXiv:2211.08466*.

- Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.
- Kunze Li and Yu Zhang. 2024. Planning first, question second: An llm-guided method for controllable question generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4715–4729.
- Yunting Liu, Shreya Bhandari, and Zachary A Pardos. 2025. Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052.
- Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 16–27.
- Kawshik Manikantan, Makarand Tapaswi, Vineet Gandhi, and Shubham Toshniwal. 2024. [Identifyme: A challenging long-context mention resolution benchmark](#). Preprint, arXiv:2411.07466.
- Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. [Multi-hop question answering](#). Preprint, arXiv:2204.09140.
- Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12:1–32.
- Edward J O’Brien, Anne E Cook, and Robert F Lorch. 2015. *Inferences during reading*. Cambridge University Press.
- Onkar Pandit and Yufang Hou. 2021. [Probing for bridging inference in transformer language models](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4153–4163, Online. Association for Computational Linguistics.
- E Margaret Perkoff, Abhidip Bhattacharyya, Jon Cai, and Jie Cao. 2023. Comparing neural question generation architectures for reading comprehension. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 556–566.
- Yin Poon, John Sie Yuen Lee, Yu Yan Lam, Wing Lam Suen, Elsie Li Chen Ong, and Samuel Kai Wah Chu. 2024. [Few-shot question generation for reading comprehension](#). In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 21–27, Bangkok, Thailand. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Manav Rathod, Tony Tu, and Katherine Stasaski. 2022. [Educational multi-question generation for reading comprehension](#). In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 216–223, Seattle, Washington. Association for Computational Linguistics.
- Ina Rösiger. 2018. [BASHI: A corpus of Wall Street Journal articles annotated with bridging links](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- John Sabatini, Jonathan Weeks, Tenaha O’Reilly, Kelly Bruce, Jonathan Steinberg, and Szu-Fu Chao. 2019. SARA Reading Components Tests, RISE forms: Technical Adequacy and Test Design. *ETS Research Report Series*, 2019(1):1–30.
- Andreas Säuberli and Simon Clematide. 2024. [Automatic generation and evaluation of reading comprehension test items with large language models](#). In *Proceedings of the 3rd Workshop on Tools and Resources for People with READING Difficulties (READI) @ LREC-COLING 2024*, pages 22–37, Torino, Italia. ELRA and ICCL.
- Franz Schmalhofer, Mark A McDaniel, and Dennis Keefe. 2002. A unified model for predictive and bridging inferences. *Discourse Processes*, 33(2):105–132.
- Murray Singer, Peter Andruslak, Paul Reisdorf, and Nancy L Black. 1992. Individual differences in bridging inference processes. *Memory & cognition*, 20(5):539–548.
- Murray Singer and Gilbert Remillard. 2004. Retrieving text inferences: Controlled and automatic influences. *Memory & Cognition*, 32(8):1223–1237.
- Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A Hearst. 2021. Automatically generating cause-and-effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170.
- Richard Thurlow and Paul van den Broek. 1997. Automaticity and inference generation during reading comprehension. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 13(2):165–181.
- Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. [Difficulty-controllable neural question generation for reading comprehension using item response theory](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 119–129, Toronto, Canada. Association for Computational Linguistics.

- Paul van den Broek, Katinka Beker, and Marja Oudega. 2015. Inference generation in text comprehension: Automatic and strategic processes in the construction of a mental representation. In Edward J. O'Brien, Anne E. Cook, and Robert F. Lorch Jr., editors, *Inferences during reading*, pages 94–121. Cambridge University Press.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic questions and where to find them: Fairytaleqa—an authentic dataset for narrative comprehension. *arXiv preprint arXiv:2203.13947*.
- Jason D Yeatman, Jasmine E Tran, Amy K Burkhardt, Wanjing Anya Ma, Jamie L Mitchell, Maya Yablonski, Liesbeth Gijbels, Carrie Townley-Flores, and Adam Richie-Halford. 2024. Development and validation of a rapid and precise online sentence reading efficiency assessment. In *Frontiers in education*, volume 9, page 1494431. Frontiers Media SA.
- Eric Zelikman, Wanjing Ma, Jasmine Tran, Diyi Yang, Jason Yeatman, and Nick Haber. 2023. [Generating and evaluating tests for k-12 students with language model simulations: A case study on sentence reading efficiency](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2205, Singapore. Association for Computational Linguistics.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

A Bridging Inference Examples

When we developed the taxonomy of bridging inference, we referred to a sample passage and a list of example questions provided from [Cain and Oakhill \(1999, p.495\)](#). Table 5 presents our analysis of the given questions based on the taxonomy.

B Annotation Guidelines

To validate the newly developed taxonomy of bridging inference questions, we annotated an in-house RC item bank. Annotation was done with regards to the text and the questions including stem, key and distractors (see details in Table 6).

C Prompts

We present examples of our few-shot prompting design for pronominal bridging (Figure 3) text-connecting (Figure 7) and gap-filling (Figure 8) respectively. The rules are identical for both the standard and CoT prompts; the only difference is that CoT includes a text hint and reasoning in the training examples (see blue highlight in the figure). Accordingly, in the CoT condition, we expect the output to include a text hint and reasoning along with the generated questions.

Reading Passage:

Debbie was going out for the afternoon with her friend Michael. By the time they got there they were very thirsty. Michael got some drink out of his duffel bag and they shared that. The orange juice was very refreshing. Debbie put on her swimming costume, but the water was too cold to paddle in, so they made sandcastles instead.


They played all afternoon and didn't notice how late it was. Then Debbie spotted the clock on the pier. If she was late for dinner, her parents would be angry. They quickly packed up their things. Debbie changed and wrapped her swimming costume in her towel. She put the bundle in her rucksack. Then they set off for home, pedalling as fast as they could. Debbie was very tired when she got home, but she was just in time for dinner.

Question	Annotation
Literal information	
Who did Debbie spend the afternoon with?	The answer is in the first sentence. There is a partial paraphrase: "going out for" vs. "spend".
Where was the clock?	The answer is in the second sentence of the second paragraph.
Text-connecting inference	
Where did Michael get the orange juice from?	This requires bridging inference: <i>drink = orange_juice</i> . This is both a referential and semantic link (hypernym: drink – hyponym: juice). Recognizing this link requires background knowledge and both components are near each other in the text.
Where did Debbie put her towel when she packed up her things?	The answer is in sentences 5–6 of the second paragraph. This involves recognizing a part-whole relationship (towel–bundle), which is an ad-hoc, situational reference.
Gap-filling inference	
Where did Debbie and Michael spend the afternoon?	One component (afternoon) is in the text, but the location (the beach) is not. It must be inferred as a plausible missing piece of the situation model.
How did Debbie and Michael travel home?	The text says "set off for home" (a paraphrase of "travel"). The mode of travel is inferred from "pedalled", enriching the situation model.

Table 5: Analysis of a reading passage and associated reading comprehension questions with inference annotations. The passage and questions are adapted from Cain and Oakhill (1999, p.495).

Dimension	Options	Note
Inference	Factual / Literal	The answer is explicitly stated in the text, exactly matching the question. No inference needed.
	Pronominal	Resolving pronouns (e.g., "Who does 'he' refer to?").
	Pronominal Bridging	Requires resolving a pronoun and using it as a cue to infer the correct answer.
	Text-Connecting	Requires connecting two explicitly stated components, typically using noun phrases.
	Gap-Filling	Involves filling in a missing but easily inferred piece of information not directly stated in the text.
	Vocabulary	Tests the reader's knowledge of word meanings.
	Other	Any other type, such as comparison or author intent.

Table 6: Annotation guidelines for the in-house item bank.



Task: Given a passage, you are going to generate text-connecting inference questions.

Follow these steps to answer the user queries.

Step 1 - find two concepts (primarily nouns or noun phrases) that are connecting AT LEAST 2 or 3 sentences, but their relationship is not explicitly stated.
Please follow the rules:

- The two concepts should not contain any same word. Incorrect example: "the Ocean" and "Pacific Ocean" share a word "Ocean". Correct example: "flowers" and "rose".
- The two concepts should only exist in two different sentences.
- The second concept should not be a pronoun that explicitly refers to the first concept.
- there are different possible subtypes of text-connecting you may find from the passage:

Subtype 1: Coreference without a pronoun nor repetition (share word): This refers to instances where two or three sentences are linked together by two noun phrases in the passage that refer to the same real-world entity. Correct examples: "boys and girls" referring to "students" from the previous sentence, "manager" referring to the "CEO" from the previous sentence. Incorrect examples: "he" referring to "John" (as "he" is a pronoun), "the show" referring to "TV show" (because this is a repetition and they share the word "show", unless there is more than one show described in the passage).

Subtype 2: Whole-to-part relation. For instance, "mom" refers to "parent", "bride" can refer to the "wedding" from the previous sentence, and "walls" can refer to the "construction project" mentioned earlier.

Subtype 3: implicit causal relation without a clue word

Subtype 4: events happen in the same time, etc.

Step 2 - based on two concepts you have identified, generate a multiple-choice question with three distractors. The question should use the relationship between the two concepts as a hint to connect information between sentences.

Step 3 - follow additional rules when writing the questions: 1) do not ask a question that requires extra background knowledge beyond this identified text-connecting relationship to answer. 2) do not ask a question that directly asking "what does XX refer to". 3) lightly paraphrase the question and option without introducing new inference. 4) do not write correct answer longer than the distractors.

Step 4 - iterate this process for 2 times to get 3 different questions. Do not force to generate more questions if you cannot find more places.

Step 5 - Output by following the exact format as examples so that it can be directly converted to csv format (do not have any title like (**questions**)).

Here are some example passages and example questions:




*****Given passage:*****
A greenhouse is a building where plants such as flowers and vegetables are grown. It usually has a glass ...

*****Examples:*****
 PassageName\Inference Type
 \Text Hint\Reasoning
 \Stem\Option 1\Option 2\Option 3\Option 4\Key

Greenhouse\text-connecting bridging

Many vegetables and flowers are grown in greenhouses in late winter and early spring, when it is still too cold to grow plants outside. Then these plants move into the soil outside as the weather warms up.


"these plants" links to "many vegetables and flowers" as a part to whole relation in the previous sentence.
When do greenhouse vegetables and flowers move into the soil outside?when the weather warms up\when heating is not working\in early spring\when there is no rain\1..

*****New Passage:*****
Parallax is the perceived change in position of an object seen from two different places ...

Figure 7: Few-shot prompting using Chain-of-Thought for generating text-connecting inference.

Task: Given a passage, you are going to generate gap filling inference questions. This question asks for a piece of information outside of the text, i.e. general knowledge, with information in the text to fill in missing details in the passage.



Follow these steps to answer the user queries.

Step 1 - Find a concept in the passage that you think general background knowledge will be required to comprehend the text. There are three possible subtypes:
 Subtype 1: two or three sentences are connected without a pronoun but by a common sense that is not stated in the passage.
 Subtype 2: infer the result from a given situation based on a stated causal relationship. for example :The passage implies that if, then _____. The result should not appear in the passage.
 Subtype 3: to give an example based on the characteristics inferred from the text. for example: Which of the following could be an example of _____. Note that the example should not appear in the passage.

Step 2 - generate a multiple-choice question with three distractors.

Step 3 - follow additional rules when writing the questions: 1) do not ask a question that can be directly answered from the passage. 2) do not ask a question that directly asking "what does XX refer to". 3) do not write correct answer longer than the distractors. 4) the distractors should be incorrect and should not be confusing.

Step 4 - iterate this process for 2 times to get 3 different questions. Do not force to generate more questions if you cannot find more places. You don't need to generate each subtype.

Step 5 - Output by following the exact format as examples so that it can be directly converted to csv format (do not have any title like (**questions**)).


Here are some example passages and example questions:


*****Given passage:*****
A greenhouse is a building where plants such as flowers and vegetables are grown. It usually has a glass ...


*****Examples:*****
 PassageName\Inference Type
 \Text Hint\Reasoning
 \Stem\Option 1\Option 2\Option 3\Option 4\Key

Greenhouse\Gap-filling
 \Also, greenhouses can get very hot from the sun's heat, so gardeners have to make sure that it does not get too hot for the plants.
 Greenhouses usually have vents that can be opened to let excess heat out. Some greenhouses have electric exhaust fans that automatically turn on if it gets too hot in the greenhouse. A greenhouse is the place for tender plants such as tomatoes, cucumbers, and aubergines.
 \Infer the result from a given situation based on a stated causal relationship

\What is likely to happen if a greenhouse fails to control the heat in summer?\The greenhouse will grow more plants.\The greenhouse will become smaller.\Tender plants inside the greenhouse will not grow well.\Less gardeners will be needed to water the plants.\3







*****New Passage:*****
Parallax is the perceived change in position of an object seen from two different places ...

Figure 8: Few-shot prompting using Chain-of-Thought for generating gap-filling inference.