



**HAL**  
open science

## GPT-3-driven pedagogical agents for training children's curious question-asking skills

Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, H el ene Sauz eon, Pierre-Yves Oudeyer

► **To cite this version:**

Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, et al.. GPT-3-driven pedagogical agents for training children's curious question-asking skills. *International Journal of Artificial Intelligence in Education*, 2023, *International Journal of Artificial Intelligence in Education*, 10.1007/s40593-023-00340-7 . hal-03890447

**HAL Id: hal-03890447**

**<https://hal.science/hal-03890447v1>**

Submitted on 16 Dec 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destin ee au d ep ot et  a la diffusion de documents scientifiques de niveau recherche, publi es ou non,  emanant des  tablissements d'enseignement et de recherche fran ais ou  trangers, des laboratoires publics ou priv es.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

# GPT-3-driven pedagogical agents for training children’s curious question-asking skills

Rania Abdelghani<sup>\*1,2</sup>, Yen-Hsiang Wang<sup>4</sup>, Xingdi Yuan<sup>3</sup>, Tong Wang<sup>3</sup>, H el ene Sauz eon<sup>†1,5</sup>, and Pierre-Yves Oudeyer<sup>‡1,3</sup>

<sup>1</sup>Inria France

<sup>2</sup>EvidenceB, France

<sup>3</sup>Microsoft research, Canada

<sup>4</sup>National Chung Hsing University, Taiwan

<sup>5</sup>University of Bordeaux, France

December 16, 2022

## Abstract

Students’ ability to ask curious questions is a crucial skill that improves their learning processes. To train this skill, previous research has used a conversational agent that propose specific cues to prompt children’s curiosity during learning. Despite showing pedagogical efficiency, this method is still limited since it relies on generating the said prompts by hand for each educational resource, which can be a very long and costly process.

In this context, we leverage the advances in the natural language processing field and explore using a large language model (GPT-3) to automate the generation of this agent’s curiosity-prompting cues to help children ask more and deeper questions. We then used this study to investigate a different curiosity-prompting behavior for the agent.

The study was conducted with 75 students aged between 9 and 10. They either interacted with a hand-crafted conversational agent that proposes ”closed” manually-extracted cues leading to predefined questions, a GPT-3-driven one that proposes the same type of cues, or a GPT-3-driven one that proposes ”open” cues that can lead to several possible questions. Results showed a similar question-asking performance between children who had the two ”closed” agents, but a significantly better one for participants with the ”open” agent.

Our first results suggest the validity of using GPT-3 to facilitate the implementation of curiosity-stimulating learning technologies. In a second

---

\*rania.abdelghani@inria.fr

†Equal supervision

‡Equal supervision

step, we also show that GPT-3 can be efficient in proposing the relevant open cues that leave children with more autonomy to express their curiosity.

**Keywords**— GPT-3, Prompt-learning, conversational agents, curiosity, question-asking, children

## 1 Introduction

Curiosity is defined as "the desire to know, to see, or to experience that motivates exploratory behavior directed towards the acquisition of new information" [34]. It is a crucial factor that modulates positively the individuals' learning experience and outcomes [55, 21]. Moreover, research shows strong evidence that curiosity is a malleable skill that can be elicited by both verbal and non-verbal cues [9, 15, 31], and by promoting comfort with uncertainty and encouraging questioning to resolve it [23].

However, and although inquiry-based learning is at the forefront of modern education [8, 29], reports show that curiosity decreases with formal schooling [12]. More particularly, curiosity-based question-asking (QA) behaviors are shown to be almost absent from today's classrooms: children's inquiries are very often low-level questions that are memory-based and that do not require much cognitive work [16, 20], which does very little to increase their curiosity [5]. Possible explanations of this phenomena can be children's tendency to overestimate their knowledge and, thus, not having the need to ask further questions [16], their incapacity to formulate syntactically-correct questions [20], their fear of their classmates' judgments [43], etc.

To remedy these problems, recent studies have proposed to take advantage of the benefits of using conversational agents in learning [3] and proposed new curiosity-prompting educational technologies. For example, authors in [2, 1] proposed interactions with conversational agents that are capable of stimulating children's questioning and exploration. Their results showed positive effects on children's curiosity-driven behaviors as well as on their domain-knowledge learning progress, thus suggesting the pedagogical efficiency of their methods.

A logical follow-up to scale-up and generalize these studies is exploring ways to automate the said conversational agents' behaviors in order to facilitate their implementation on a larger scale and for different learning tasks. More particularly, we think of the natural language processing (NLP) field and the large language models (LLMs) that showed an impressive ability in predicting word sequences that resemble the way people write( [13], [61]...).

In this context, we study using the recent LLM GPT-3 to implement conversational agents that can prompt children's curiosity about a given text-based educational content, by proposing some specific cues. For this, we use GPT-3 to replicate the exact same curiosity-prompting strategies that were validated in [1] and generated by hand. We then investigate the validity of this automation method by comparing its impact on children's divergent QA skills with respect to the hand-crafted condition. In a second step, we explore using GPT-3 to propose a new curiosity-prompting

behavior for our agent that aims to better support the children’s needs of competence, autonomy and relatedness during the QA training. Indeed, following the self-determination theory [48, 17], these are fundamental factors to promote for more efficient educational trainings.

Our study contributes both to the ongoing research on how to design effective curiosity-prompting learning technologies in general and on innovative ways to use NLP methods in educational settings.

## 2 Related work

### 2.1 Epistemic curiosity and QA training in education

Curiosity plays an important role in learning from infancy [53] to adulthood [45], and in fostering academic achievement, particularly for children with low socio-economic status [50]. It can be seen as a personality trait ([28]), a psycho-emotional state aroused by external situations([39], [27]),[4],...) or as a malleable skill that can be trained through teaching specific practices. Indeed, several theories and frameworks such as Lowenstein’s ”Information gap theory” ([37] or Murayama’s process account of curiosity in learning ([26]), suggest that enhancing one’s curiosity-driven learning should focus on training the awareness of missing information and the appropriate inquiry strategies to compensate for it.

One of the most important inquiry strategies to be trained is question-asking: research such as in [2] suggest that the more curious students are by trait, the more divergent-thinking questions they will ask (i.e., high-level questions that require making hypotheses, predictions, ideas inferences... [14]). To assess this kind of behavior, studies such as [58] presented a question-asking based exploration game and showed that the precision of the questions that children generated was a significant predictor for their curiosity, as reported by their teachers. On another hand, Abdelghani et al., [1] investigated using conversational agents to train children’s divergent QA skills by proposing specific hand-generated semantic and linguistic cues. The study then investigated whether such trainings can help children engage in intrinsically-motivated explorations of their learning environments and gain new knowledge on their own. Their results showed, indeed, positive effects of the intervention on children’s curiosity-driven behaviors and their subsequent domain-knowledge learning progress. However, the agents presented in these tools had very low intelligence level (i.e. guided by pre-defined automatons) and restrained children to think about predefined specific questions which can limit their usefulness, usability and transfer in different classroom activities.

### 2.2 Natural language processing and conversational agents for building educational applications

Advances in the NLP field have helped create several opportunities to build innovative and high-quality educational applications. Work in this area has facilitated the semantic, syntactic and grammatical analysis of

students' typed and spoken inputs and the detection of their potential errors. Another interesting use of NLP in education is the automation of tasks that are traditionally executed by hand such as creating the curriculum and the evaluation materials content. Indeed, authors in [56] for instance, showed that NLP-based methods can be used to extract the key scientific ideas and concepts from science educational resources that, according to human experts, are important for a better and deeper understanding of the domain. More importantly to this study, NLP-based methods have also shown to be efficient in the question-generation tasks (i.e. generating questions from educational resources) which can help teachers alleviate the time-consuming process of creating the assessment and evaluation materials ([6], [54], ..).

On a final note, it is also important to say that earlier research has showed several positive pedagogical impacts of using NLP-based social conversational agents on children's perception of their own competencies and their attitudes towards learning ([51]).

### 2.3 GPT-3 and prompt-based learning

Over the past few years, the NLP field has been focusing on the use of pre-trained large language models (LLMs) as they demonstrate impressive powers in generating natural language and are easier to use in downstream tasks ([11, 7]). Indeed, LLMs are trained on enormous amounts of text data, giving them the ability to learn novel tasks only by taking a few examples as context.

However, adapting a pre-trained LLM to the context of a specific downstream task is still a quite complex task as it mostly relies on fine-tuning: a learning method that requires performing a further training of the pre-trained parameters of the LLM using labelled data that is specific to the task in question. This method can be problematic for several reasons: it can be extremely costly and hard to perform for super large models that have an enormous number of parameters [44], data can be scarce in many cases, the model is fine-tuned separately for each different task, etc. As an alternative, methods such as prefix-tuning have emerged [33], where only a small percentage of the model's parameters are updated for task-adaptation. We also find adapter-tuning methods [19] that only add a few trainable parameters per new task, allowing a high degree of parameter sharing. These simpler techniques have achieved comparable results to fine-tuning but still require machine learning expertise for their implementation, which can be a major brake for using them in educational settings where most stakeholders do not have this expertise.

In this context, GPT-3, one of the recent and most performing LLMs, came along with an even simpler technique called "prompt-based learning": it is a new way to handle a wide range of NLP tasks by leveraging the LLM's ability to model natural language. It consists of verbalizing the task to be completed in natural text (this is what is called a "prompt"), with no update of the model's parameters. It is also possible to provide examples in the "prompt" to show the model how the task should be completed; this is known as K-shot learning, where K refers to the number of examples provided. This technique has been applied in a wide range

of applications such as knowledge-based question-answering [22] but, up to our knowledge has not yet been applied for curiosity-based questions generation. In addition, prompt-based learning with GPT-3 has achieved very strong performances in zero-shot and one-shot settings (i.e. with only zero or one labelled example), without any fine-tuning [7]. This is an enormous advantage for this study as it allows to implement a simple-to-use technology that can be easily usable and/or adapted for a different context by the education community (i.e. teachers, caregivers ..) without having to have any expertise in machine learning.

## 2.4 Current study

This study comes as a follow-up for previous work that showed promising results when using a hand-crafted conversational agent to prompt children’s curiosity, using specific strategies [2, 1]. However, and despite their pedagogical relevance in enhancing QA skills, these previous implementations presented important drawbacks, mainly concerning the difficulty of their use on a larger scale and in different contexts, given that the agents had low computational powers. The prompting approach was also arguable as it left children with very little choice over the questions they had to formulate since the agent had pre-defined behaviors.

In the mean time, progress in the NLP field and its integration in the educational applications is becoming more and more solid thanks to impressive performances in tasks such as question-asking, question-answering, key concepts-extraction, etc. In this context, our current study tries to leverage these new advances to address the two main drawbacks of the previous work mentioned above. Indeed, we test automating the implementation of the validated curiosity-prompting approaches by using the GPT-3 LLM. In a further step, we also investigate using GPT-3 to endow our conversational agent with a new divergent QA-prompting strategy that aims to provide children with more autonomy in the task and help support better their curiosity. Thus, with our study we aim to make two main contributions :

- Investigate the validity of using a LLM (GPT-3) as a tool to facilitate the implementation of QA trainings on a larger scale. For this, we take previous human-generated curiosity-prompting approaches as a gold standard and try to replicate them using GPT-3. We then compare the effect of the two implementations on children’s divergent QA performances.
- Investigate using a LLM (GPT-3) to propose a new behavior for the agent in order to implement a student-directed curiosity-prompting approach that can lead to formulate divergent questions in a more autonomous and free setting. Here, we test providing children with a more open type of curiosity-prompting cues and investigate their divergent QA performances compared to our previous settings.

### 3 Study design

”Kids Ask” is a prototype for a web-based educational platform that involves an interaction between a child and a conversational agent during reading-comprehension tasks. The platform was designed for our previous study ([1]) and offers three work spaces: one to train children’s ability to self-reflect on their knowledge (workspace 1), train their question-asking skills (workspace 2) and train their follow-up exploratory behaviors (workspace 3).

For our current study, and since we are only focusing on QA training via raising knowledge-gap awareness, we choose to reuse the same two first work spaces. We only adapt the agent’s implementation in the second workspace: it is driven by utterances that were created manually for the previous study and driven by GPT-3 for the current one; see more details of this aspect in the Technical Implementation Section 3.3.

Concretely, the current ”Kids Ask” implementation offers the two following work spaces:

- **Workspace to elicit curiosity via training self-reflection:** This space contains a domain-knowledge quiz; the items can be skipped if children decide that they don’t know the answer. If, however, they submit an answer, the agent asks them to report a confidence level in this answer with a 5-Likert scale (from ”Super not confident” to ”Super confident”). This strategy is inspired from studies such as [16] where authors show that children do not engage in curiosity-driven behaviors mainly because they fail to see the uncertainties they have about their own knowledge. We therefore use the confidence report strategy to help them reflect on their own knowledge and think deeper about what they do and do not know. In addition, work in [46] suggests that giving children the possibility to skip a question allows them to report low certainty which, in its turn, can motivate curiosity-driven behaviors. The rationale behind this session is to poke children’s curiosity by encouraging them to report low certainty levels. Indeed, and as explained above [4], detecting uncertainty is the primary fuel for curiosity-driven behaviors such as question-asking.
- **Workspace to train question-asking skills (The QA training space):** In this space, children have different texts relating to a theme of their choice: they choose one of the six themes that they have worked on in the first workspace. For each text, they interact with the agent that will try to help them think of divergent questions by giving them specific linguistic and semantic cues. These cues are a questioning word that is either combined with 1) an answer to a specific question about the text, that was generated by hand, 2) an answer to a specific question about the text, that was generated by GPT-3, or 3) important key words that are related to the text and that were generated by GPT-3. See the Experimental Conditions Section 3.1 for more details about the composition of these cues.

The agent chooses to give one of these three types of cues depending on the condition the child is assigned to: hand-crafted incentive

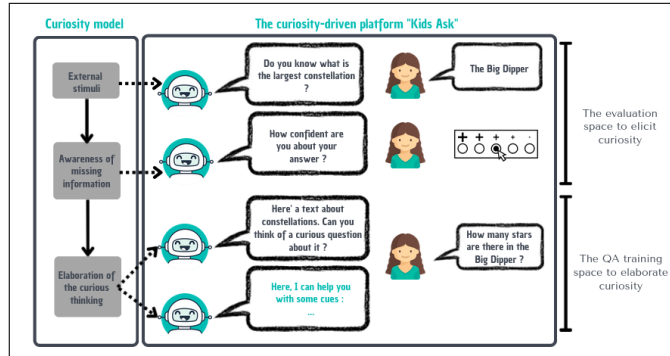


Figure 1: Illustration of the agent’s strategies in the different work spaces of ”Kids Ask”, with relation to our curiosity model

agent, GPT-3-driven incentive agent or GPT-3-driven open agent. For the three conditions, and similarly to the previous study in [1], the agent does not give any feedback concerning the question formulated by the child.

Participants started with the general quiz relating to 6 themes using the first workspace. They then chose their favourite topic and moved to the QA-training space where they had reading tasks related to their topic of choice.

See Figure 1 for a general illustration of the ”Kids Ask” platform.

### 3.1 Experimental conditions

We set up three experimental conditions :

1. Hand-crafted incentive agent : A control group where the agent is controlled by human-generated incentive cues that lead to predefined specific divergent-thinking questions, i.e., high-level questions that require making hypotheses, predictions, ideas inferences... [14] (Group 1).
2. Automated incentive agent : A first test group where the agent is controlled by GPT-3-generated incentive cues that lead to predefined specific divergent-thinking questions (Group 2).
3. Automated open agent : A second test group where the agent is controlled by GPT-3-generated open cues that lead to several possible divergent-thinking questions about the text (Group 3).

More precisely, the cues for the first two groups had the exact same type and consist of the following: a questioning word + an answer to one possible divergent question. Combined, these two cues are meant to lead children to think of one specific curious question about the text. **Example:** For a text about Louis Pasteur, the agent gives the cues 'What difference' and 'The vaccine avoids the disease whilst medicine



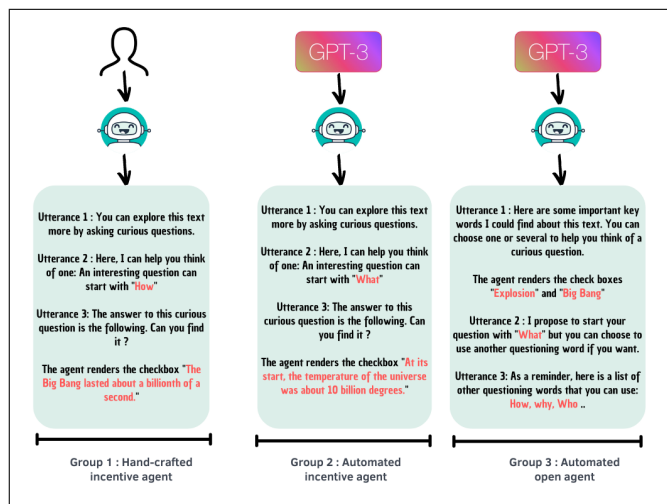


Figure 2: Difference between the agents' behaviors and implementations for the three experimental conditions

treats it' in order to lead the child to ask the question 'What is the difference between vaccines and medicine?'. This question is considered to be divergent as its answer is not initially available in the text and requires the child to link two ideas [14]. For the Group 1, these cues were chosen by researchers and teachers as they were judged to be leading to important questions for children to ask. For the Group 2, these cues were generated by GPT-3 (see details about the technical implementation in section 3.3.

For the third group, the agent proposes two key words that are related to the main ideas of the text. These words constitute cues that can lead towards several possible curious questions about the text. To be in the same conditions as the other groups, we also give a possible questioning word. But because the possibilities are multiple in this condition, the agent tells children that they can use different starters if they want to and reminds them of the most commonly-used ones. **Example:** For a text about Louis Pasteur, the agent gives the cues **What other** and **"Canning, Freezing"**. Multiple divergent questions are possible with these words: **"What is the difference between canning and freezing?"**, **"Why is Pasteurisation better than freezing?"**, **"What are the other ways to preserve food?"**....

See Section3.2.2 for the rationale behind the choice of these cues, Figure 2 for details about the difference between the agents' dialogues and implementations for the different conditions and Section4.2 for snips of the agents' behaviors in "Kids Ask".

## 3.2 Design rationale

### 3.2.1 Choice of the interaction workflow

As mentioned above, we choose to start our interaction with a skippable general knowledge quiz in order to elicit children’s curiosity about one or different topics that we propose. Since this quiz contains items relating to different topics, it also serves us as a tool to help children choose the topic they want to work on during the next task, in the QA training space. Indeed, giving children the choice over their tasks has shown to have a strong positive effect not only on their motivation and engagement but also on the amount of knowledge they can learn and their perceived competence [10].

### 3.2.2 Choice of the cues proposed by the agent

For the QA-training space, we focus on divergent questions in general as they require higher-level thinking and are more related to epistemic curiosity ([14, 49]). In order to guide children towards this type of questioning, we choose specific cues for our agents to propose (as described in 3.1). This choice is based on three main ideas:

1. The first type, i.e. the answers to specific divergent questions that the agents gave to Group 1 and 2, are motivated by the results of our two previous studies [1, 2]: when generated by hand, this type of cues showed to be rather efficient in training children’s divergent QA skills. However, and as mentioned above, this manual implementation, being very time-costly, makes the generalization and scale-up of the approach, even though pedagogically efficient, very difficult. Therefore, and in order to facilitate this agent implementation, we use GPT-3 to generate the exact same type of cues. We then compare children’s performances when having the cues generated by humans vs. by GPT-3 (i.e. Group 1 vs. Group 2) in order to investigate the validity of using GPT-3 for this task.
2. The motivation behind the second type of semantic cues (i.e. the open key words that were given to Group 3) is to propose a prompting strategy that leaves children with more autonomy and choice over the questions they can formulate during the training. Indeed, having heavily teacher-directed activities that impose the same specific pieces of information to all children (like the case for our first type of cues) is likely to result in giving trainings that do not match with children’s different competency levels, i.e., their zones of proximal development (ZPD). Indeed, guiding children to ask specific questions can lead them to ask questions that are too complicated for them or, on the contrary, questions that they already know the answer to. This can mean that the training is likely not to support the users’ curiosity since this latter is tightly related to each individual’s ZPD [40]. Indeed, learners decide whether or not to engage in pursuing a piece of information based on if they perceive that this latter will be in their ZPD or not; which differs among learners depending on their expertise for the subject matter at hand [40]. On the other

hand, the non-matching tasks can alter children’s interest in them as as this latter is driven by choice that, at its turn, depends on the learner’s expertise and ZPD [52]. We also think that it is important to give different choices in order to avoid the cases where teachers’ or researchers’ choice of the semantic cues is heavily influenced by their preconceptions and beliefs about children’s knowledge gaps [25, 59]. And finally, this lack of control over the task might also alter with children’s sense of autonomy and competency, which are central factors for motivation and curiosity according to the Self-Determination Theory( [48], [17]).

Therefore, and inspired by studies such as [56] that show the efficiency of using NLP methods to extract the core concepts and/or ideas from scientific educational resources, we propose to explore using GPT-3 to extract relevant keywords that can lead to asking divergent questions about a given text. These keywords can then be proposed by the agent to raise children’s awareness about the important concepts of the texts that need further investigations, and lead them into thinking of different open questions. Investigating children’s performance when having these GPT-3-generated specific vs. open cues (i.e. Group 2 vs. Group 3) could therefore help us investigate different contexts where children’s curiosity can be induced more efficiently.

3. Finally, we choose to maintain the linguistic help (i.e. the questioning word that can be used to start the question) for our three conditions as question-generation tasks can be quite difficult for children, as shown in [16].

### 3.2.3 Choice of the technology

In order to automate our conversational agents’ behaviors, we choose to work with the pre-trained LLM GPT-3 and to use prompt-based methods<sup>1</sup>. First, we choose to work with GPT-3 given that it has significantly more parameters and was trained upon text corpus containing a wider variety of topics, compared to its predecessor GPT-2 for example. Moreover, it has demonstrated robust performances in various downstream tasks, and more importantly to this study, in knowledge-based question-asking and question-answering tasks [62]. Indeed, such results motivate a follow-up work to move towards more complex types of questions like high-level and divergent questions.

Second, and with the aim to implement an easy-to-use system that can be accessible for the broad teaching community, we choose to work with GPT-3 in a prompt-based methods setting ([7]). Indeed, and as discussed above, prompt-based learning consists simply of giving pieces of natural text to the model as an instruction in order to adjust it to the specificities of the task in question and generate the output needed (in our case, this output is a set of cues as described in the Conditions

---

<sup>1</sup>It is to be noted that in this article we use the word "prompt" in two different contexts : (1) prompts we give GPT-3 to generate the cues proposed by the agent and (2) prompts we give to children to help them generate divergent questions.

section 3.1). So with this simple implementation, we avoid methods such as fine-tuning that require a machine learning expertise and collecting a large set of specific data, etc. Instead, we provide a system that can be easily used by practitioners (e.g. teachers) and/or adapted to different activities of their choice (e.g. change the language, target a different type of questions ..), without the need for a specific expertise. Finally, investigating these strategies can also be advantageous for the educational technologies community as it will help show new ways of using GPT-3 for building curiosity-prompting educational applications.

### 3.3 Technical implementation

#### 3.3.1 Generating the agents' cues using chain-of-thoughts prompting in GPT-3

For our experiments, we prompt GPT-3 in a zero-shot setting [7], meaning that the model predicts an answer based only on a context and natural language description of the task, i.e. a prompt, without giving it any examples. Concretely, we provide a data-set containing the text-based educational resources: short scientific articles with a mean of 109 words per article, that will serve the model as context. Our task consists of building the proper prompts that can lead GPT-3 to produce a questioning word combined with either a sentence (for the group 2) or a pair of related keywords (for the group 3) that can be used as cues to formulate a divergent question relevant to the context. See Figure 3 for details about the prompts approaches we retained.

Before deciding on which prompting approach to choose, we run several trials where the following configurations were tested:

- Story then instruction vs. instruction then Story: At the very beginning of the development, we did try a set of different ways to form the prompts, including the ordering of story vs instruction, using special indicators to highlight story/instruction etc. and prompt formats (either to have a new line after the indicator). We do not observe significant difference from these variants, probably because LLM at the scale of GPT-3 is sufficiently robust on this format dimension. We decide to stick with a format (as shown in Figure 3) that is considerably straightforward and readable so practitioners do not need to specially design the prompts.
- Zero-shot vs. One-shot vs. Few-shots vs Fine-tuning: In all the experiments described in this work, we prompt GPT-3 in a zero-shot manner. The reason is two-fold. 1) There are many recent works in the NLP community report that pre-trained LLMs such as GPT-3 can be sensitive to the specific data points provided as few-shot examples [30, 38, 63, 41]. For instance, LLMs are observed to suffer from “recency bias”, i.e., they overly rely on the examples that are located closer to the end of the prompt and thus tend to bias the output towards copying from the most recent examples. Despite many works have been proposed to alleviate such issues, it remains to be solved to have a few-shot prompting strategy that is

[Story]:  
 The bees perform a dance when they want to indicate to their colleagues the position of a nectar source. They perform a figure eight choreography on the vertical combs of the hive. If the bee moves up and down the figure of eight, it means that it should face the sun; if it moves in the opposite direction, it means that it should face away from the sun. Bees also communicate by means of "vibrations": the "grooming" vibration, for example, which a bee emits when it needs a colleague to clean it, or the "beggar" vibration. The bees also do a semi-circular dance.

[Instruction]:  
 For the above [Story] , generate 5 [Key ideas].

[Key Ideas]:

1. Bees use dance to communicate the position of nectar sources.
2. Bees communicate by means of vibrations.
3. The figure eight choreography is used to communicate the position of the nectar source.
4. The semi-circular dance is used to communicate the direction in which the bee should face.
5. The "grooming" vibration is used to communicate the need for cleaning.

[Instruction]:  
 For each of the [Key Ideas] above, generate 2 related words on the same topic.

[Related Words]: // Used as the semantic cues for the automated open agent

1. Bees: honey, pollination
2. Vibrations: communication, signals
3. Figure eight: choreography, dance
4. Semi-circular: dance, communication
5. Grooming: cleaning, hygiene

[Instruction]: Use the above [Related Words] to ask 5 divergent questions.

[Divergent Questions]: // The questioning words are used as linguistic cues for the two automated agents

1. What other ways do bees communicate?
2. What other functions do vibrations serve in bee communication?
3. What are other examples of figure eight choreography in nature?
4. What are other purposes of the semi-circular dance?
5. What are the consequences of not grooming?

[Instruction]:  
 Answer the above [Divergent Questions] using sentences with less than 10 words.

[Answers]: //Used as the semantic cues for the automated incentive agent

1. Bees also communicate by smells and tastes.
2. Vibrations also serve to warn of predators.
3. Other examples of figure eight choreography include birds flying in formation and fish swimming in schools.
4. The semi-circular dance can also be used to communicate the direction in which the bee should fly.
5. Not grooming can lead to infection and disease.

Figure 3: Prompting approaches used to control the agents behaviours

generally working better. 2) We want to emphasize that in most real-world scenarios such as pedagogical applications, practitioners are less likely to have machine learning expertise nor sufficient computational resources to fine-tune an LLM, or to optimize the LLM outputs by providing the model with different combinations of example. We envision that the proposed system can have greater accessibility if it does not require too much expertise from the users. Lastly, we want to mention that our method is orthogonal and applicable to bleeding-edge prompt-tuning methods from the NLP community [47, 38, 36].

In all experiments, we use the text-davinci-002 variant of GPT-3 and a temperature of 0.7 during the prompting process.

### 3.3.2 Agent behavior

As we have no control over GPT-3’s outputs, we choose not to put children in direct contact with the model in order to avoid any potential offensive interactions. To do so, we opt for an offline implementation like the following:

- We use GPT-3 to generate the agents’ cues like described in the section above.
- For more transparency of our proposed method, we repeat the procedure several times for the same resources and we take some of the outputs randomly as cues to appear in the platform (even if they lead to convergent questions).
- After verification that these cues do not include any inappropriate or offensive words, we translate them into French using DeepL and include them in a dedicated database. It is to be noted that during all of our experiments, we did not see any offensive output (see more details in the results section 7.1.1).

The agent is then connected to this database containing the different educational resources and the cues for each one of them, for the three conditions. The cues for the hand-crafted agent are included manually in the database. Depending on the child’s assigned condition, the agent composes the dialogue utterances in order to include the appropriate support. We changed the utterances between the questions to avoid repetition in the agent’s dialogue: **Example** : If, for the first question, the agent says **”Here are two cues to help you think of a curious question about the text, can you use them ?”** than for the following one it will choose another replica such as **”Can you combine these two cues to generate a curious question about the text ?”** [24].

## 4 Methods

### 4.1 Participants

We recruited 75 4th grade students belonging to three classes from two French primary schools, they were between 9 and 10.5 years old and assigned either to the group 1 with the incentive manual agent (24 with 13 boys and 11 girls), group 2 with the incentive GPT-3 agent (26 with 13 boys and 13 girls) or to group 3 with the open GPT-3 agent (25 with 11 boys and 14 girls).

The groups were assigned with a pseudo-randomized method after collecting profile data regarding the age, the device use frequency, the curiosity-related measures, the reading score and their domain-knowledge scores (see 5 for more details of the measures). As shown in Table 1, we had three balanced groups that were not different in terms of these initial profile measures, with the exception of the familiarity with using digital tools measure. But since the study was run in a controlled environment where researchers could help participants when they have any technical difficulties, we assume that this difference does not have a significant effect

of our measure of interest, i.e. formulating divergent questions. To ensure this, we run an ANCOVA analysis in order to investigate the effect of the participants’ exposure to digital tools score on their divergent QA performance, between the three groups. Results showed that the interaction was the same between the three groups suggesting that the exposure variable does not affect the equilibrium between our participants (p-value=0.23).

Table 1: Profile measures for the three conditions

Measure	Group 1	Group 2	Group 3	p-values
Age	9.4 ± 0.43	9.56 ± 0.38	9.52 ± 0.44	0.31
Device use frequency	27 ± 7.66	31.39 ± 6.12	32.57 ± 5.28	0.008*
Curiosity trait	27.2 ± 4.38	29.69 ± 4.53	28.36 ± 4.75	0.16
Perception of curiosity	37.54 ± 6.72	38.3 ± 6.66	33.92 ± 9.12	0.09
Reading ability	279.55 ± 93.02	297.37 ± 104.61	293.19 ± 73.6	0.77
Ability to ask div. questions	1.33 ± 0.8	2 ± 1.17	2.2 ± 0.93	0.11
Domain-knowledge quiz score	6.9 ± 4.6	8.77 ± 3.58	8.72 ± 4.05	0.23

## 4.2 Procedure

The study consisted of three sessions of 1h within the same week: one for collecting the profile measures and pre-intervention measures mentioned above, one for the divergent-QA training and one for the post-intervention assessment. See study timeline in Figure 4 and details of the different measures in the data collection paragraph 5.

Our study was approved by the institute’s ethics committee (certificate n°2019-23) and started after having all participants parents’ signed consents.

**Session 1:** During the first session, we presented the study to participants and collected the age and gender data. We also administered the questionnaires concerning the curiosity trait, the perception of curiosity and device use frequency. We then run the reading abilities test and the question-asking fluency test (QA fluency test), see more details of the tests in the section 5. We also took time to explain what divergent questioning means and highlighted the difference between divergent and convergent questions. Participants then went to the evaluation space of ”Kids Ask” and began the interaction with a general domain-knowledge quiz relating to six themes. For each question, children had the choice to either skip the item, by clicking on the ’I don’t know, I want to skip this question’ button, or answer it. If they do choose to answer, they are asked to report their confidence level in their answer: they had a 5-Likert scale from ’Super not confident’ to ’Super confident’ (see Figure 5).

Once finished, participants chose their favourite topic among the ones they had during the quiz. They were told to choose carefully as this is the topic they will be working on during the next task.

**Session 2:** During the QA training session, participants were asked to read short texts relating to the theme they chose. They then started

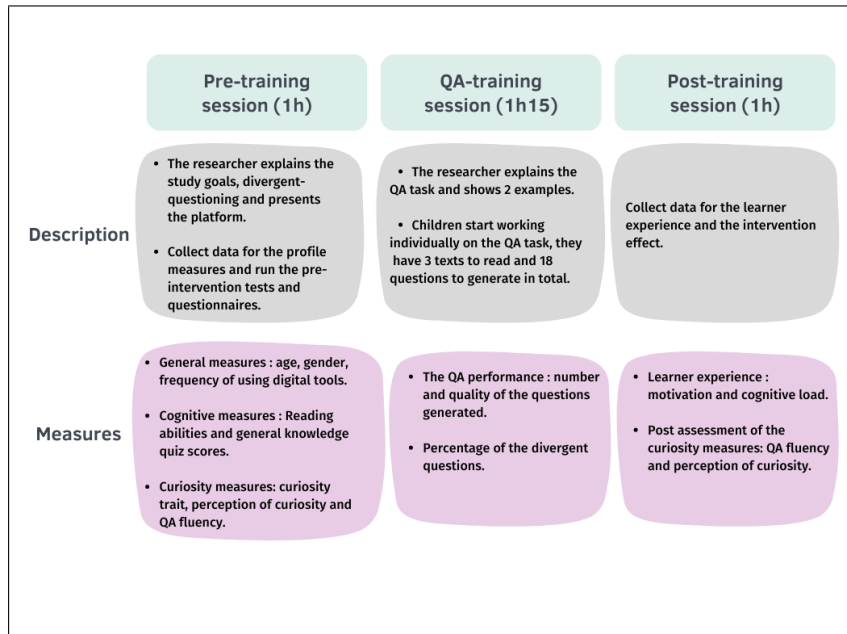


Figure 4: Study timeline and measures

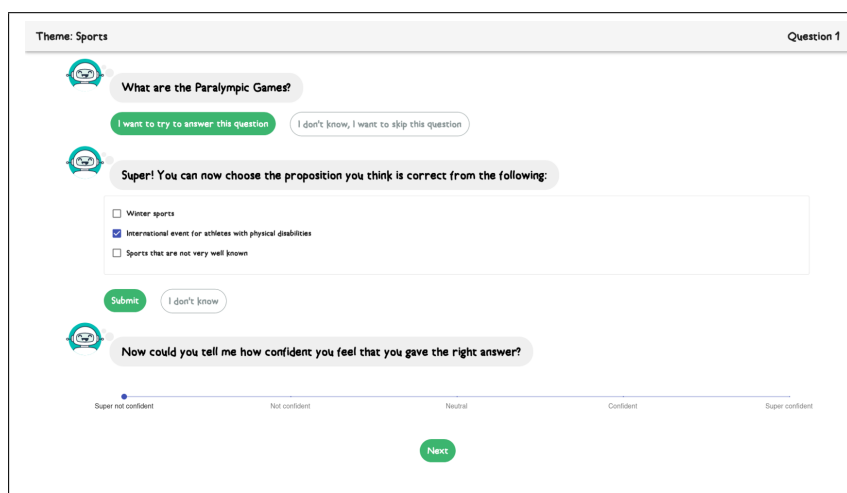


Figure 5: Interface for the domain knowledge quiz



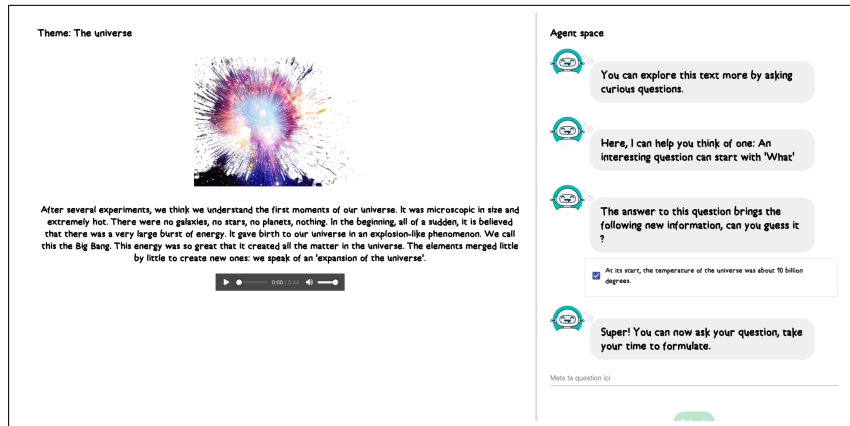


Figure 6: Agent behavior for the incentive conditions

working on generating divergent-thinking questions relating to these texts using the cues provided by the agent. They also had audio players for each text to help them if they had reading difficulties. The texts were selected from online resources and children magazines (Sciences et vie Junior and Quelle Histoire) and were edited in order for them all to have six sentences and an average number of 109 words per text.

To begin working on a text, participants were asked to read or listen to it and then click on the 'I finished reading' button once they understand it. This button enables the 'discussion' with the agent, in the agent's space on the right section of the screen. As explained above, the agent helps the child generate divergent-thinking questions about the text by giving specific linguistic and semantic cues. The two incentive agents (hand-crafted and GPT-3-driven) had the same behavior; their only difference lied in the generation method of the cues presented. While the "open" had slightly different utterances to go with the type of cues it proposes. See Figures 6 and 7 for the difference between the conditions.

During this session, children had to process three texts and generate six questions per text with the agent, making a total of 18 questions. They were not restricted in terms of time, apart from the session length. They had the application running on tablets and worked individually.

**Session 3:** Participants answered the post-intervention surveys for the general motivation, types of motivation and the task load. They also retook the QA fluency test and the curiosity perception questionnaire they had pre-intervention (see measures timeline in 4 and further descriptions in the "Data collection and instruments" section 5).

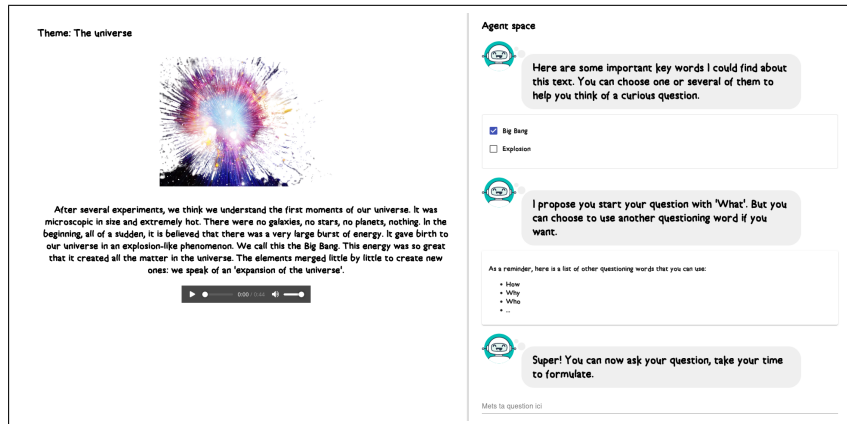


Figure 7: Agent behavior for the "open" condition

## 5 Data collection and measures

### 5.1 Children profile measures

This includes general measures: age, gender and the exposure to digital tools, the curiosity trait ([35]). We also have three cognitive measures: the reading abilities using the standardized test in [32], the domain-knowledge quiz score and the divergent QA fluency test : the ability to ask divergent questions. This latter is computed as the number of divergent questions that children could generate after reading a short article offline, within 2 minutes. This same divergent QA fluency test was used in previous studies in [1, 2].

### 5.2 Cues-generation performance

#### 5.2.1 Semantic relatedness

In scoring the quality of the cues generated, we evaluate their semantic relatedness to the text in question. This is an important measure given that the goal is to lead children to ask divergent questions that are still closely related to the educational resource. For this we use a 5-scale Likert and run a human annotation; the items the grid we used can be found in the AppendixA.

#### 5.2.2 Divergence level

We also take into account the divergence level of the cues, compared to the text's context. This will help us understand better children's divergent QA performance later on, because it gives us a clue about whether or not the support we're giving children is pushing them towards divergent thinking in the first place. We also use human annotation to assess this aspect, following the grid described in B.

## 5.3 Question-asking performance

### 5.3.1 Percentage of divergent questions generated

This measure consists of counting the total number of the correct questions generated during the training and computing the percentage of the divergent questions amongst them. If a question is repeated, it is only taken into account once. Questions that did not use the agents' cues were still taken into account if they were still relevant to the general context of the text under discussion. See more details about our acceptance criteria in the appendix C.

Once a question is accepted, we evaluate whether or not it is divergent by checking if its answer is explicitly stated in the text. This is based on Gallagher's classification [14] where he posits that divergent questions require children to make hypotheses, ideas linking etc. **Example:** The question "What was the temperature of the universe at its beginning?" is considered to be divergent, whereas "What caused the birth of the universe?" is considered convergent as the answer to it is explicitly stated in the text.

All data was anonymized: the coder could only see the identifiers that children were given randomly at the beginning of the intervention. The inter-rater reliability was of 87.6%.

### 5.3.2 Syntactic scores for the questions generated

In scoring the quality of the questions generated, we used the grid that emerged from the classification in [14] and computed a question's score as the sum of the openness of the question, its syntactic construction and its use of the questioning words. The grid scores a question from 2 to 8 points (see details in appendix D).

For each participant, we calculated an average score for all the questions generated during the training. This annotation process was performed by the two researchers who led the experiments in schools, they had an inter-rater reliability of 76.8%. For data points where we saw difference between the two coders, we used an average score.

## 5.4 Curiosity measures

### 5.4.1 Divergent QA fluency

In order to have a behavioral measure for curiosity pre- and post-intervention, We run an offline test where we give participants a short text to read then ask them to write as many questions about it as they can, within 2 minutes. We then compute the divergent QA fluency as the sum of divergent questions that children were able to generate during this fixed period of time.

### 5.4.2 Perception of curiosity

To assess children's initial perception of curiosity, we use Post's [43] validated CIAC questionnaire. The 24-item questionnaire is designed to evaluate how much children relate social matters and epistemic questions to

curiosity. It also evaluates children’s attitude towards epistemic curiosity: Personal inclination, Social relevance, negative opinion, fear of classmates negative judgement; self-efficiency in asking questions.

## 5.5 Learning experience

### 5.5.1 Motivation measures

We use Vallerand’s scale to assess the intrinsic and extrinsic motivational mechanisms [60]. The scale is composed of three sub-scales that differentiate: intrinsic motivation (possible scores from 0 to 9 points), extrinsic motivation (possible scores from 0 to 9 points) and amotivation (possible scores from 0 to 3 points). All questionnaire items are yes or no questions.

### 5.5.2 Cognitive load

The workload of the task, as perceived by the participants was assessed using the NASA-TLX workload multi-dimensional scale developed in [18]. Information about the intensity of six workload-related factors are used in order to estimate a reliable measure of workload: mental demand, physical demand, temporal demand, performance, effort and frustration.

## 6 Research goals and hypotheses

- The GPT-3-driven incentive agent will lead children to similar curiosity-driven question-asking performances as the hand-crafted one.
- The open agent, compared to the two incentive agents, will lead children to ask more curiosity-driven questions.
- The divergent question-asking performance will be a strong positive predictor for children’s curiosity traits, as reported by their parents.
- The QA training, irrespective of the agent’s implementation, will enhance children’s perception of curiosity.
- The QA training, irrespective of the agent’s implementation, will enhance children’s ability to ask divergent questions.

## 7 Results

Before moving to study children’s interaction with ”Kids Ask”, we start by evaluating GPT-3’s performance in generating semantic and linguistic support that can prompt curiosity. This evaluation is important to help us better understand the system’s impact on children later on.

### 7.1 GPT-3 performance

#### 7.1.1 General observations

As a first step, we start by verifying GPT-3’s outputs with regards to two aspects : offensiveness and the general relatedness to the context of

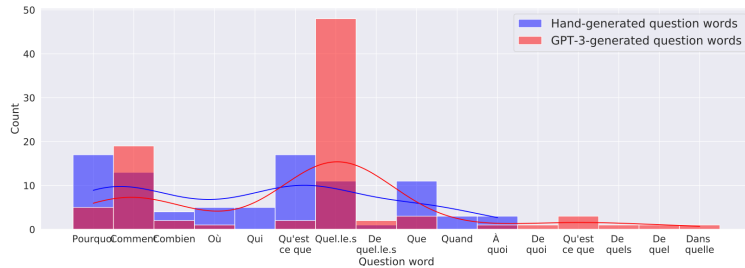


Figure 8: **Generating the linguistic cues : Distribution of the question words proposed in the "Kids Ask" platform.**

the educational resource in hand. For the offensiveness measure we run a human annotation with a 5-Likert scale (from "Not at all offensive for a 10-year-old child" to "very offensive for a 10-year-old child"). Results showed that 100% of the data generated was scored as no at all offensive ( $m=5$  ;  $SD=0$ ). We then evaluated this output depending on its relevance to the general context of the text (e.g. for a text about the Big Bang, data-points about the universe in general are considered relevant, however output about different topics like the sports ... are considered not relevant). Here again, we run a human annotation and saw that 100% of our generated data was annotated as relevant to the task.

### 7.1.2 Generating the linguistic cues

In order to investigate GPT-3's performance in generating the questioning words, we proceed by comparing its output with what was produced by humans wrt the variety of the cues proposed and their complexity level.

The variety of the question words proposed is an important observation for us in order to ensure that children do not receive the same words to use during the whole training. To assess this, we start by plotting the histogram of the cues generated by hand and by GPT-3 (See Figure8). We then investigate the similarity between the variances of the two distributions. We use a Levene's test given that the question words generated by hand deviated from normality. Results showed no support to reject similarity between the two conditions ( $p\text{-value}=0.65$ ).

Finally, we also compared the complexity levels of these cues. Indeed, and as suggested in [20], compound questioning words (e.g. "What difference", "What if" ..) are more likely to lead towards higher-level questions than simple words like "What", "Where" etc. For this, we compare the proportions of occurrence of this type of questioning words when generated by hand vs. by GPT-3. The 2-sample z-test showed no support to reject the hypothesis of having two similar proportions ( $p\text{-value}=0.51$ ).

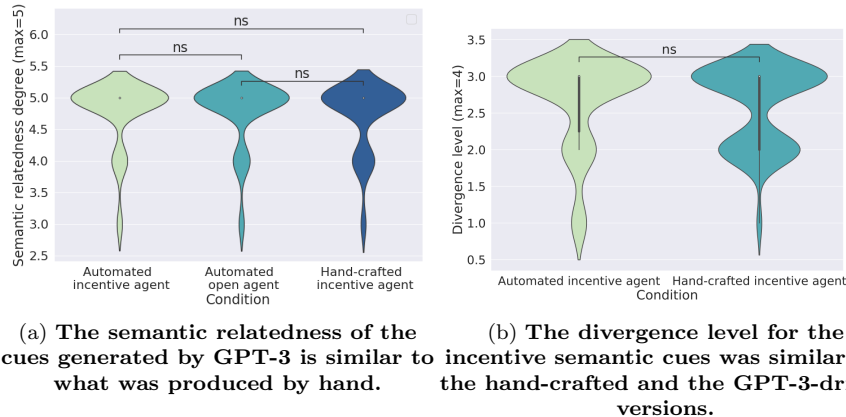


Figure 9: Generating the semantic cues : Evaluation with respect to the semantic relatedness to the context and the divergence level.

### 7.1.3 Generating the semantic cues

Here again, we base our evaluation on comparing the semantic cues generated by GPT-3 for the incentive and the open agents to those generated by hand. We start by comparing the semantic relatedness degree to the context of the proposed educational resources, as described in the Measures Section 5.2. As seen in Figure 9, we see no significant difference between the three conditions for this measure. The one-way ANOVA test failed to exclude similarities between at least two groups ( $F(2,74)=0.41$ ,  $p\text{-value} = 0.66$ ), with a power test of 0.98 indicating that we only have 2% chance to mistakenly conclude that none exists. Assumptions about the normality of our distributions, homogeneity of variances and independence between the observations were confirmed pre-running the test.

We also compare the divergence levels of the proposed cues for the two incentive agents. It makes sense to assess this aspect given that the incentive agents propose specific cues that will heavily condition children’s questions. Thus, if the proposed cues are already mentioned in the text i.e., convergent, they will automatically lead children to generate convergent questions. However, this is not the case for the open agent’s cues, i.e., the keywords, as the questions’ divergence levels here do not rely on the words themselves but on how children choose to use them. For this reason, we only assess the cues’ divergence levels for the two incentive agents, by running a human annotation following the grid described in B. As it can be seen in Figure 9, we run a t-test and saw the two distributions did not differ ( $t=0.9$ ,  $p\text{-value}=0.37$  and  $\text{power}=0.14$ ).

Similarly to the linguistic cues, here we also see results suggesting that our prompting approaches were successful in generating divergent semantic cues that are closely related to the context in question and are comparable to what we had in the hand-crafted condition.

## 7.2 Children’s performance

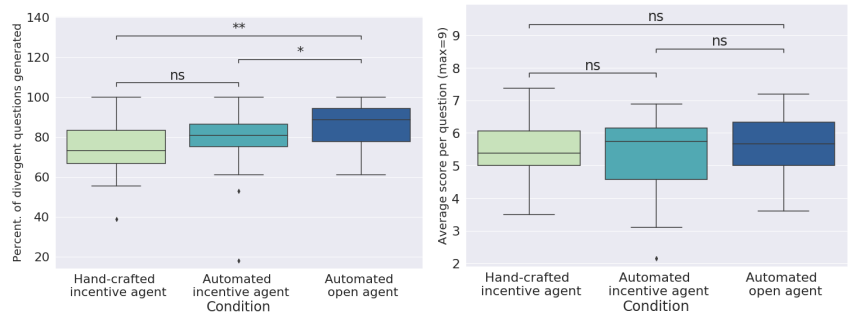
Moving on to studying the effect of our system on children, we start by looking into their QA performance during the training. We then investigate the relationship of this performance with the curiosity trait scores and evaluate the learning experience in terms of its pleasantness and its perceived cognitive load. We finally report the mid-term effect of this interaction on children’s QA fluency pre- and post-interaction using offline tests.

Before running this analysis, we verified that the imbalance in children’s familiarity with using digital tools among the three groups did not affect our different measures. Indeed, we run several ANCOVA tests with, at each time, one measure related to their interaction with ”Kids Ask” as a dependent variable, and the exposure to digital tools score as a covariate. All of our tests concluded that this score did not affect our different measures.

### 7.2.1 Divergent question-asking performance

In a first step, we start by evaluating the general usefulness and accessibility of the cues proposed by the agents by seeing how often children actually chose to use them to generate their questions. Concretely, this amounts to counting the number of questions formulated by children to which the answers coincide with the proposed semantic cues for the incentive agents, and the number of questions that use one or both of the keywords proposed for the open agent. Results showed no difference between the two incentive agents for the percentage of questions where they used the agents’ cues ( $M=77.28\%$  and  $SD=23.08$  for the hand-crafted agent;  $M=76.12\%$  and  $SD=23.88$  for the automated agent with a p-value for the T-test of 0.86 and a power test of 0.05). However, children with the open automated condition used the agent’s cues more often ( $M=91.78\%$ ,  $SD=8.83$ ) and were significantly different from the hand-crafted incentive agent (p-value= 0.03, power=0.85) and the GPT-3-driven one (p-value=0.005 and power=0.81); suggesting that the keywords-based prompting strategy may have been more accessible for children.

Moving on, we analyze the effects of the three different agents on children’s divergent QA performances by comparing the percentages of the divergent questions generated. These percentages are computed as explained in Section 5.3.1. In a first step, we perform a one-way ANOVA test between the two groups; results show indeed a significant difference in the performances ( $F(2,72)=4.11$ , p-value=0.02). We then proceed to perform T tests in pairs : results showed no significant difference between the two incentive agents, i.e. the hand-crafted one and the GPT-3-driven one ( $t=0.88$ , p-value=0.38 and power=0.14) suggesting that children’s performances were similar, with only 14% chance to see a difference between the two. However, and as expected, the group that interacted with the automated open agent had a significantly better performance than the Hand-crafted incentive agent ( $t=-3.17$ , p-value=0.003 and power=0.82) and the GPT-3-driven incentive agent ( $t=-1.88$ , p=0.04 and power=0.45); see Sub-Figure (a) of Figure 10. These results were rather expected as the



(a) **Participants with the incentive agents (hand-crafted or GPT-3-driven) had similar divergent QA performances. Those who had the automated open agent had a significantly better performance than the two incentive groups.**

(b) **Participants with the three different conditions had a similar average syntactic score per question they generated.**

Figure 10: Question-Asking performances during the training, using the QA-training workspace

open agent leaves more space for children to imagine their own different curious questions, rather than restrain them to a specific one that they may or may not be able to imagine (like it is the case for the two incentive agents). The normality and the homogeneity of variances of our data were verified pre-running our tests.

Finally, we also investigated the quality of the questions generated, following the standardized grid described in 5.3.2. These scores give us an idea about the syntactic and semantic quality of the children’s questions. Results showed no significant difference between the two incentive agents ( $t=0.57$ ,  $p\text{-value}=0.57$  and  $\text{power}=0.08$ ). The same observation was also found between two automated agents, i.e. incentive vs open ( $t=-0.81$ ,  $p\text{-value}=0.41$  and  $\text{power}=0.12$ ) and between the control agent (the hand-crafted incentive agent) and the the automated open one:  $t=-0.29$ ,  $p\text{-value}=0.77$  and  $\text{power}=0.06$ . See Sub-Figure (b) of Figure10.

Taken together, our first results suggest the validity of the GPT-3 prompting approach we used to generate answers that lead children to ask curious questions. Indeed, the GPT-3-driven incentive agent, compared to the human-driven one, led to similar performances in terms of quantity and quality of the questions children generated, as well as a similar dependency on the semantic cues. In a second step, our results suggested that giving related keywords that were generated by GPT-3 may be a stronger strategy to help stimulate children’s curiosity-driven questioning. Indeed, compared to the first two conditions, children chose to use the agent’s words more frequently to generate their questions. They also ended up formulating more divergent questions while maintaining similar syntactic scores even though they had no semantic formulations for their cues and reported no additional cognitive load.





Figure 11: For the automated open agent, children’s curiosity trait scores were strong predictors for their divergent question-asking abilities.

### 7.2.2 Relationship with the curiosity trait as reported by parents

In order to study the interaction between children’s divergent QA performance, their curiosity trait scores (as reported by their parents) and their assigned conditions, we performed an ANCOVA test between the three groups, with the percentage of divergent questions as a dependent variable and the curiosity trait score as a co-variate. Results indicate a significant interaction between the divergent-QA performance and the curiosity scores, within the three conditions ( $F(2,71) = 4.06$ ,  $p\text{-value}=0.02$ ). Pairwise comparisons are then conducted in order to identify groups that are statistically different : we apply Bonferroni’s multiple test correction. The post-hoc analysis show a statistically-significant difference only for the open agent condition ( $p\text{-value}=0.006$ ). We also tested the significance of differences between the Pearson correlation coefficients for the three groups, by applying a Ronald Fisher z-transformation. As it can be seen in Figure 11 that the correlations were similar for the two incentive agents ( $z=1.7$ ,  $p\text{-value}=0.14$ ); however, the open agent led to a significantly different relationship ( $z=1.7$ ,  $p\text{-value}=0.04$ ) between the hand-crafted incentive agent and the automated one and ( $z=2.8$ ;  $p\text{-value}=0.002$ ) between the automated incentive agent and the automated open one.

These results can posit that our open agent is more associated with curious thinking given that it leaves more space to children to formulate questions of their own choice and thus, to express their own curiosity: in this condition, the more children were curious by trait, the more divergent questions they were likely to generate (see Figure11). However, in the case of the two incentive agents, we saw no significant correlations between their performance in the task and their curiosity trait. This can be explained by the idea that, given that children were restrained to think of pre-defined specific questions, their task was more of a social curiosity training rather than an idiosyncratically one, as explained in Section3.2.2.

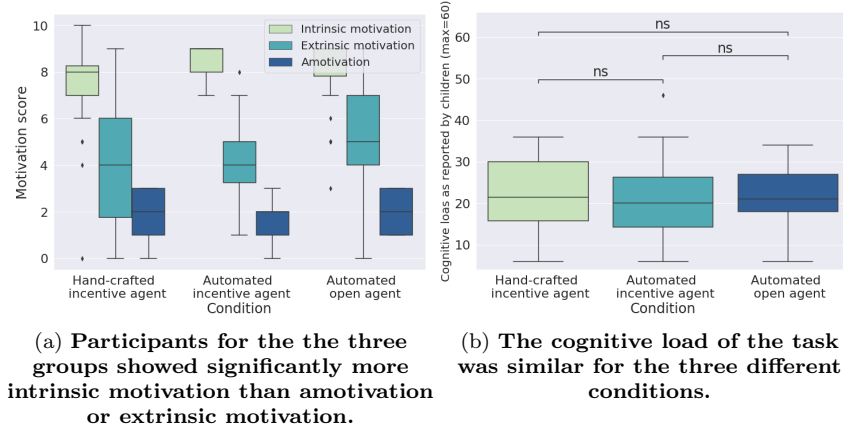


Figure 12: Learner experience measures

## 7.3 Learning experience

### 7.3.1 Motivation measures

We spotted no difference between the children’s intrinsic motivation during the training with the three different agents. However, we saw that the three groups were significantly more intrinsically than extrinsically motivated to do the task ( $t=3.3$ ,  $p<0.001$ ,  $\text{power}=0.98$  for the hand-crafted incentive agent,  $t=12.02$ ,  $p<0.001$ ,  $\text{power}=1$  for the automated incentive agent and  $t=3.9$ ,  $p<0.001$ ,  $\text{power}=0.97$  for the automated open agent); see sub-figure (a) in Figure12.

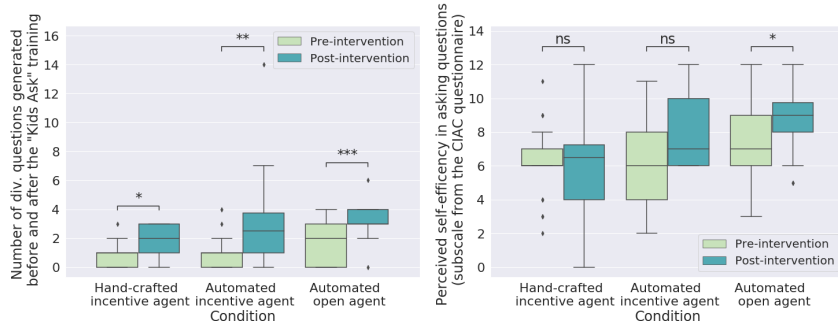
### 7.3.2 Cognitive load

On a final note, we also investigated the cognitive load as perceived by children, following the Nasa-tlx scale as described in Section 5.5. The aim here is twofold: 1) evaluate whether or not children perceived that the answers provided by GPT-3 were harder to assess than those provided by humans. 2) see if the lack of a semantic formulation in the open cues for the open automated agent makes the question-generation task harder for children. To address these goals, we perform pairwise t-tests. Results showed indeed no significant differences between the two incentive agents ( $t=-0.4$ ,  $p=0.1$ ,  $\text{power}=0.06$ ) or between the incentive and open agent ( $t=-0.3$ ,  $p\text{-value}=0.7$ ,  $\text{power}=0.06$ ).

## 7.4 Intervention effect: pre-post measures

### 7.4.1 On the divergent QA fluency

Children’s ability to ask divergent questions is an important aspect of their epistemic curiosity and can be seen as a relevant behavioral indicator for it. For this reason, we evaluate the efficiency of our curiosity



(a) Participants from the three conditions were able to improve their self-efficacy changed more positively divergent QA abilities after the "Kids with the intervention for those who Ask" interaction, as shown by the divergent QA fluency test pre- and post-training. (b) Children's perception of their QA self-efficacy changed more positively divergent QA abilities after the "Kids with the intervention for those who Ask" interaction, as shown by the divergent QA fluency test pre- and post-training.

Figure 13: Pre- Post-intervention curiosity-related measures

training by comparing children's divergent QA-fluency scores pre- and post-intervention, between the three conditions. This measure is collected during offline fluency tests like described in section 5.4.2).

For this, we performed a two-way repeated measures ANOVA test in order to investigate the impact of the time on the divergent QA performance between the three conditions. Results showed a statistically significant effect of the time ( $F(2,72)=46.95$ ,  $p\text{-value}<0.001$ ) and the condition ( $F(2,72)=3.9$ ,  $p\text{-value}=0.002$ ), with no effect of the interaction condition:time on the performances ( $F(2,72)=2.23$ ,  $p=0.11$ ); see Sub-figure (a) in Figure 13. These results show that participants from the different groups benefited from the curiosity intervention, suggesting the general validity of the prompting approaches used and the benefits of using conversational agents to stimulate children's divergent question-asking. Hypotheses concerning the absence of outliers and normality of the divergent QA-fluency scores were verified pre-running the test.

#### 7.4.2 On the perception of curiosity

In analyzing the effect of our intervention on children's attitude towards epistemic curiosity, we also run a two-way mixed ANOVA test to investigate the difference in the CIAC scores ([43]) before and after our intervention, between the different groups. Results show a non-significant interaction between time and the condition type ( $F(2,72)=3.87$ ,  $p\text{-value}=0.06$ ) but reveal a significant change in the scores for all participants pre- and post-intervention (time Effect:  $F(1,72)=15.74$ ,  $p\text{-value}<0.0001$ ). This suggests the general efficiency of our proposed method in changing children's perception of curiosity in the classroom.

Furthermore, we were interested in investigating the Self-Efficiency

sub-scale of the CIAC questionnaire which measures how children perceive their own skills in asking questions. Indeed, one important aspect for this study’s effectiveness is its efficiency in helping participants observe and be aware of the progress they can achieve with such trainings in order to motivate them to use them more during learning. As shown the sub-figure (b) in Figure 13, the three groups had better scores for the post-intervention test and the interaction between time and the groups was significant ( $F(2,72)=0.038$ ,  $p\text{-value}=0.04$ ). The post-hoc pairwise tests showed a significant increase in this score only for the group that interacted with the open GPT-3 driven agent ( $p=0.014$ ). This result aligns well with what’s discussed in the design rationale section 3.2.2 where we suggest that giving children tasks with a larger margin of choice and autonomy can effect their perception of their self-efficiency.

## 8 Discussion

With this work, we investigate using prompt-learning with the pre-trained LLM GPT-3 to implement different conversational agents that aim to stimulate children’s curiosity during learning, using some specific verbal cues. We validate our approach by using human annotations to score these cues and by comparing their impact on children’s divergent QA performances with human-based interventions.

With our implementation, we tried to address transfer learning in educational settings in general and in curiosity-prompting, more particularly. We also address the questions concerning performing the question-generation and question-answering NLP tasks in languages other than English. Indeed, before being proposed to children, our GPT-3-generated curiosity-prompting cues went through an automatic translation step into French, using DeepL. Adding this translation layer still led us to rather encouraging general results wrt the quality of the prompts generated and their positive impacts on children’s divergent question-asking performances. Such findings can lead us to suggest that our approach can be fairly transferable to other languages where NLP-based educational systems are still quite limited.

Moving forward, we saw that our system was more efficient in prompting children’s curious question-asking when we had an open agent that proposed keywords leading to several possible questions, rather than the incentive agents that led to one specific and pre-defined question. These observations were somehow expected given that the open agent was designed with the intention to propose a student-directed training. Indeed, we hypothesize that such a setting will lead children to work on questions that correspond to their expertise level which will result in better perception of the self-efficiency in the task [40] and therefore, can lead to better engagement and to even more learning progress [42, 57]. Furthermore, we saw similar syntactic scores for the three different groups, as well as similar scores for the tasks’ perceived cognitive load. This can be explained by the observation that children did not use the agent’s semantic cues exactly as they were formulated to generate their questions. Instead, they extract the important ideas and used them to formulate a question on

their own. This observation may also reinforce the validity of our choice to use keywords for the third group.

In analyzing the link between the curiosity trait scores and the performance during the divergent QA training, we saw a strong correlation only for the group that interacted with the open agent. Several ideas can be posited to explain these findings : first, it is possible that children who are more curious by trait, since they are more familiar with asking questions without any syntactic support, found in the open agent a more favorable condition to express their curiosity. Second, it is possible that giving children QA activities that they can transpose to their ZPDs helped their curiosity. We lean more towards the second explanation since we saw a significant increase in children’s perception of their QA skills. Thus, in this condition, children’s ability to ask divergent questions can potentially be seen as a positive predictor for their curiosity trait. This can be suggested as we see a significant effect of the agent’s condition on the interaction between children’s divergent QA fluency progress and their reported curiosity traits. This is an interesting finding as it addresses the question of proposing new behavioral measures for curiosity that are suited to e-learning environments, rather than using the classic self-report measures.

Finally, we saw high, yet similar motivation scores for the three conditions. This result was also somehow expected given the novelty aspect of our approach in general (i.e. using tablets in the classroom, interacting with a conversational agent, the QA task itself, ..), and also, given the short length of the training. Indeed, we can speculate that the motivation scores for the incentive conditions where children have little control over the task, would drop if we implement a longer training. And therefore, expect to see more intrinsic motivation for the open agent group where the task offers more degrees of choice and autonomy.

Given the positive results we found with the open agent condition, we can afford to propose it as a promising prompting approach to train children’s curiosity. It also allows us to think of new possible implementations where our agents facilitate the curious QA task by guiding children to identify the key concepts/words of a resource first, then encouraging them to use these words to formulate a related curious question. Indeed, we can imagine that decomposing the task in this way can make it easier for children to learn and to re-use in the future. Such an approach can also be adopted by teachers in classes without the use of any technology.

## 9 Limitations and future directions

One relevant drawback to our current implementations is the lack of feedback we give children about their performance during the task. Indeed, and even though we see enhanced perceptions of the self-efficiency, we do not know if this is directly connected to their perception of their performance or to other factors, like the repeatability of the exercise, bias in the self-report measure, etc. Therefore, one future direction for this work will be to explore ways to use LLMs in general in order to analyze children’s questions and give them immediate feedback about their relevance, their

divergence level and their syntactic construction. Recent work such in [62] has explored using prompt-learning methods to evaluate the generation of convergent questions and showed rather encouraging results. Therefore, one possible track can be to take inspiration from their methods and explore adapting them to divergent questions.

Another factor to be considered while evaluating this training is its short duration in time. Indeed, children were only asked to do this task during one short session which may have facilitated their engagement and played a role in the positive results we saw. Therefore, one future direction for this work is to propose a longer version of our training and investigate its long-term impact on children’s curiosity-driven behaviors.

Finally, adding more participants and more educational resources related to different knowledge components could reinforce the statistical power of our tests and therefore, foster the generalization of our findings. For example, a future track can aim to study implementing our approach with school activities where the pedagogical material is not purely text-based (e.g. includes images, diagrams, etc).

## 10 Conclusion

In this work, we contribute to the promotion of using NLP methods in building curiosity-prompting educational technologies. We show that using simple methods such as prompt-based learning with a LLM can be a valid option to implement artificial learning peers that can guide children into asking relevant and curious questions. Our work and results motivate the implementation of such curiosity-prompting approaches both in normal classroom settings and e-learning environments.

## Acknowledgements

This work has been funded by the educational technologies start-up EvidenceB and the French National Association of Research and Technology (ANRT). The authors thank the teachers who participated in this study and the research team members who helped conduct the experiments in classes.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

## A Grid for the semantic relatedness measure

Likert scale used to annotate the semantic relatedness of our cues with respect to the text in question :

- 1 point if the cue is not at all related to the text’s context
- 2 points if the cue is not related to the text’s context
- 3 points if the cue is somehow related to the text’s context
- 4 points if the cue is related to the text’s context
- 5 points if the cue is super related to the text’s context

## B Grid for the divergence level measure

Likert scale used to annotate the divergence level of our cues with respect to the text in question :

- 1 point if the cue is explicitly stated in the text
- 2 points if the cue is not explicitly stated but can be implied from the text
- 3 points if the cue is not at all stated in the text

## C Criteria for accepting children’s questions

A question is considered correct and taken into account if:

- It is a question — and not a statement.
- It is related to the text being studied.
- It is not repeated more than one time.

Questions that did not use the agents’ cues were still taken into account if they were still relevant to the general context of the text under discussion.

**Example:** For a text about the Big Bang, a linguistic cue ‘**What**’, a semantic incentive cue “**At its start, the universe’s temperature was about 10 billion degrees**” or an open semantic cue “**Big Bang, explosion**”, we accepted questions such as:

- “What was the temperature of the universe at its beginning?” for the incentive agent or “What caused the Big Bang explosion” for the open agent: both questions use the agents propositions and are related to the text.
- “How does an explosion occur?” : used one of the agent’s keywords and changed the questioning word but is still related to the text.
- “What does ‘microscopic’ mean?”: For both agents, the question here does not use the cues proposed, but is highly relevant to the text.

However, data such as the following was not accepted:

- “It is the temperature when the universe began.” : this is a statement and not a question.
- “What is a robot composed of?” : this is not related to the Big Bang text.
- “What are dinosaurs?” : this is not a serious attempt.

## D Syntactic scores for evaluating children’s questions

In assessing the questions syntactic quality, we use the grid inspired from Aschner’s classification in [14]. The grid calculates a question’s score like the following:

- One point if the question is high-level: the answer to this question is not a simple fact (example: ‘How big is a dinosaur?’) but requires to explain a mechanism, a relationship etc (example: ‘Why were dinosaurs so big?’).
- From 1 to 4 points, based on the syntactic construction of the question :
  - 1 point for a ‘closed’ or declarative question (example: “Dinosaurs were big?”).
  - 2 points for questions with questioning words in the middle of the sentence (example: “The dinosaurs were how big?”).
  - 3 points for a question without an interrogative formulation (example: “Why the dinosaurs are big?”).
  - 4 points for a questioning word in the beginning of the sentence that has interrogative syntax (example: “Why are dinosaurs big?”).
- From 1 to 3 points, based on the use of questioning words :
  - 1 point for a declarative question i.e., with no questioning word (example: “Dinosaurs were big?”).
  - 2 points for questions with ‘Is/Are’ (example: “Are dinosaurs big?”).
  - 3 points for the use of proper questioning words (example: “How big were the dinosaurs?”).

## References

- [1] R. Abdelghani, P.-Y. Oudeyer, E. Law, C. de Vulpillières, and H. Sauzéon. Conversational agents for training curiosity-driven learning in children. *International Journal of Human-Computer studies*, 167, 2022.
- [2] M. Alaimi, E. Law, K. D. Pantasdo, P.-Y. Oudeyer, and H. Sauzeon. Pedagogical agents for fostering question-asking skills in children. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.
- [3] V. Aleven, K. R. Koedinger, and K. Cross. Tutoring answer explanation fosters learning with understanding. In *Artificial Intelligence in Education*, 1999.
- [4] D. Berlyne. A theory of human curiosity. *British journal of psychology*, 45(3):180–191, August 1954.



- [5] R. Bjork. Creating desirable difficulties to enhance learning. *Car-marthen: Crown House Publishing*, 2017.
- [6] J. Brown, G. A. Frishkoff, and M. Eskénazi. Automatic question generation for vocabulary assessment. In *HLT*, 2005.
- [7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.
- [8] J. S. Bruner. The act of discovery. *Harvard Educational Review*, 31:21–32, January 1961.
- [9] J. Ceha, J. Goh, C. McDonald, D. Kulić, E. Law, N. Chhibber, and P.-Y. Oudeyer. Expression of curiosity in social robots: Design, perception, and effects on behaviour. *CHI Conference on Human Factors in Computing Systems (CHI2019)*, page 1–12, 2019.
- [10] D. Cordova and M. Lepper. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88:715–730, 12 1996.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [12] S. Engel. Children’s need to know: Curiosity in schools. *Harvard educational review*, 81(4):625—645, December 2011.
- [13] L. Floridi and M. Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.
- [14] J. J. Gallagher and M. J. Aschner. A preliminary report on analyses of classroom interaction. *Merrill-Palmer Quarterly*, 9:183–194, 1963.
- [15] G. Gordon, C. Breazeal, and S. Engel. Can children catch curiosity from a social robot? *Proceedings of the 2015 ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*, 2015:91–98, 2015.
- [16] A. C. Graesser and N. K. Person. Question asking during tutoring. *American Educational Research Journal*, 31:104–137, 1994.
- [17] F. Guay. Applying self-determination theory to education: Regulations types, psychological needs, and autonomy supporting behaviors. *Canadian Journal of School Psychology*, 37(1):75–92, 2022.
- [18] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

- [19] R. He, L. Liu, H. Ye, Q. Tan, B. Ding, L. Cheng, J.-W. Low, L. Bing, and L. Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation, 2021.
- [20] J. Humphries and M. Ness. Beyond who, what, where, when, why, and how: Preparing students to generate questions in the age of common core standards. *Journal of Research in Childhood Education*, 29:551–561, 2015.
- [21] M. Jepma, R. Verdonschot, H. van Steenbergen, S. Rombouts, and S. Nieuwenhuis. Neural mechanisms underlying the induction and relief of perceptual curiosity. *Frontiers in Behavioral Neuroscience*, 6, 2012.
- [22] Z. Jiang, J. Araki, H. Ding, and G. Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.
- [23] J. Jirout, V. Vitiello, and S. Zumbrohn. *Curiosity in schools*, pages 243–266. 07 2018.
- [24] A. Jones, S. Bull, and G. Castellano. “i know that now, i’m going to learn this next” promoting self-regulated learning with a robotic tutor. *International Journal of Social Robotics*, 10:439–454, 2018.
- [25] J. T. Jost, A. W. Kruglanski, and T. O. Nelson. Social metacognition: An expansionist review. *Personality and Social Psychology Review*, 2(2):137–154, 1998. PMID: 15647141.
- [26] M. K., , F. L., and S. M. Process account of curiosity and interest: A reward-learning perspective. *Educational Psychology Review*, 31:875–895, Dec. 2019.
- [27] M. Kang, M. Hsu, I. Krajbich, G. Loewenstein, S. M. McClure, J. Wang, and C. Camerer. The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological science*, 20(8):963–973, August 2009.
- [28] T. Kashdan, P. Rose, and F. Fincham. Curiosity and exploration: Facilitating positive subjective experiences and personal growth opportunities. *Journal of Personality Assessment*, 82(3):291–305, June 2004.
- [29] C. C. Kuhlthau, L. K. Maniotes, and A. K. Caspari. *Guided inquiry: Learning in the 21st century*. London: Libraries Unlimited, 2007.
- [30] S. Kumar and P. P. Talukdar. Reordering examples helps during priming-based few-shot learning. *CoRR*, abs/2106.01751, 2021.
- [31] E. Law, M. Yin, J. Goh, K. Chen, M. A. Terry, and K. Z. Gajos. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 ACM CHI Conference on Human Factors in Computing Systems (CHI ’16)*, pages 4098–4110, 2016.
- [32] P. Lefavrais. *Alouette-r: Test d’analyse de la vitesse en lecture à partir d’un texte*. Paris, Centre de psychologie appliqué, 2005.

- [33] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, Aug. 2021. Association for Computational Linguistics.
- [34] J. Litman. Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition & Emotion*, 19(6):793–814, September 2005.
- [35] J. A. Litman and C. D. Spielberger. Measuring epistemic curiosity and its diversive and specific components. *Journal of Personality Assessment*, 80(1):75–86, 2003. PMID: 12584070.
- [36] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics.
- [37] G. Loewenstein. The Psychology of Curiosity: A Review and Reinterpretation. *Psychological Bulletin*, 116(1):75–98, 1994.
- [38] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [39] H. Mehta, R. Dubey, and T. Lombrozo. Your liking is my curiosity: a social popularity intervention to induce curiosity. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, pages 756–761, 2018.
- [40] J. Metcalfe, B. L. Schwartz, and T. S. Eich. Epistemic curiosity and the region of proximal learning. *Current Opinion in Behavioral Sciences*, 35:40–47, 2020. Curiosity (Explore vs Exploit).
- [41] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.
- [42] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007.
- [43] T. Post and J. H. Walma van der Molen. Development and validation of a questionnaire to measure primary school children’s images of and attitudes towards curiosity (the ciac questionnaire). *Motivation and Emotion*, 42,1:159–178, 2015.
- [44] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer

- learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [45] T. Reio and A. Wiswell. Field investigation of the relationship among adult curiosity, workplace learning, and job performance. *Human Resource Development Quarterly*, 11(1):5–30, March 2000.
- [46] C. Roebers, N. von der Linden, and P. Howie. Favourable and unfavourable conditions for children’s confidence judgments. *British Journal of Developmental Psychology*, 25:109–134, 2007.
- [47] O. Rubin, J. Herzig, and J. Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021.
- [48] R. M. Ryan and E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being, 2000.
- [49] M. Scardamalia and C. Bereiter. Text-based and knowledge based questioning by children. *Cognition and Instruction*, 9:177–199, 1992.
- [50] P. Shah, H. Weeks, B. Richards, and N. Kaciroti. Early childhood curiosity and kindergarten reading and math academic achievement. *Pediatric Research*, 84:380–386, April 2018.
- [51] A. Silvervarg Flycht-Eriksson and A. Jönsson. Towards a conversational pedagogical agent capable of affecting attitudes and self-efficacy. In *Proceedings of the Second Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity*, 2010.
- [52] L. K. Son and J. Metcalfe. Metacognitive and control strategies in study-time allocation. *Journal of experimental psychology. Learning, memory, and cognition*, 26 1:204–21, 2000.
- [53] A. Stahl and L. Feigenson. Observing the unexpected enhances infants’ learning and exploration. *Science*, 348(6230):91–94, April 2015.
- [54] T. Steuer, A. Filighera, T. Meuser, and C. Rensing. I do not understand what i cannot define: Automatic question generation with pedagogically-driven content selection. *ArXiv*, abs/2110.04123, 2021.
- [55] S. Stumm, B. Hell, and T. Chamorro-Premuzic. The hungry mind – intellectual curiosity is the third pillar of academic performance. *Perspectives on Psychological Science*, 6:574–588, 10 2011.
- [56] M. A. Sultan, S. Bethard, and T. Sumner. Towards automatic identification of core concepts in educational resources. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 379–388, 12 2014.
- [57] A. Ten, P. Kaushik, P.-Y. Oudeyer, and J. Gottlieb. Humans monitor learning progress in curiosity-driven exploration. *Nature Communications*, 12:5972, 10 2021.
- [58] N. Tor and G. Gordon. Digital interactive quantitative curiosity assessment tool: Questions worlds. *International Journal of Information and Education Technology*, 10:614–621, 2020.

- [59] J. G. Tullis and A. S. Benjamin. Cueing others' memories. *Memory & cognition*, 143(4):634–646, 2015.
- [60] R. J. Vallerand, M. R. Blais, N. M. Brière, and L. G. Pelletier. Construction et validation de l'échelle de motivation en éducation (eme) [construction and validation of the motivation toward education scale]. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, 21:323–349, 1989.
- [61] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- [62] X. Yuan, T. Wang, Y.-H. Wang, E. Fine, R. Abdelghani, P. Lucas, H. Sauz on, and P.-Y. Oudeyer. Selecting better samples from pre-trained llms: A case study on question generation, 2022.
- [63] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.