The Wizard of Curiosities:
Enriching Dialogues with Fun Facts

Frederico Vicente, Rafael Ferreira, David Semedo, João Magalhães
Universidade NOVA de Lisboa
NOVA LINCS
Lisbon, Portugal
`fm.vicente@campus.fct.unl.pt`, `rah.ferreira@campus.fct.unl.pt`,
`df.semedo@campus.fct.unl.pt`, `jm.magalhaes@fct.unl.pt`

Abstract

Introducing curiosities in a conversation is a way to teach something new to the person in a pleasant and enjoyable way. Enriching dialogues with contextualized curiosities can improve the users' perception of a dialog system and their overall user experience. In this paper, we introduce a set of curated curiosities, targeting dialogues in the cooking and DIY domains. In particular, we use real human-agent conversations collected in the context of the Amazon Alexa TaskBot challenge, a multimodal and multi-turn conversational setting. According to an A/B test with over 1000 conversations, curiosities not only increase user engagement, but provide an average relative rating improvement of 9.7%.

1 Introduction

The concept of curiosity has for decades been debated by neuroscientists and psychologists. According to Kidd and Hayden (2015), it can be framed into two research views: (1) curiosity as a natural impulse for seeking extended cognition; and (2) a phenomenon related to exploring, playing, learning, and the desire for information. Berlyne (1966) went even further, meditating about how humans had inherently a special type of curiosity, an epistemic curiosity, meaning that above the exploration and information-seeking need, humans also strive for knowledge.

Multimodal conversational task assistants (Gottardi et al., 2022) seek to guide users in accomplishing complex tasks (*e.g*. "Cooking a Strawberry Pie" or "Fixing a broken chair"), in an objective, concise, and engaging manner. Naturally, conversations are rich in knowledge and senses, that are transmitted to users in a dosed manner, towards a successful completion of the task, such that at all phases, knowledge complexity is managed. From the user's perspective, executing a task can be cognitively demanding, potentially involving learning new procedures, using new tools, and following complex task instructions. Thus, conversational assistants should not only ensure a smooth completion of the tasks but also seek to make the task execution a pleasant and entertaining experience that appeals to human senses and curiosity (Dean et al., 2020). To that end, we propose to enrich conversational task assistants with contextualized fun facts, exploiting humans' curiosity-driven information-seeking traits (Kidd and Hayden, 2015). As seen in the work of Konrád et al. (2021), trivia facts have a positive impact on conversations with virtual agents, if used correctly. Hence, in this paper, when dialoguing about a complex task, the user is guided through a sequence of steps as shown in the example in appendix A. Any attempt to fruitfully extend a conversation flow must be done with care. Thus, dialog curiosities should be used as a dialog-enriching element that seeks to maximize user satisfaction. User's psychological factors aligned with the agent efficacy and correctness will be determining aspects. Inspired by Berlyne (1966)'s work, and by the computational model of curiosity of Wu et al. (2012), we propose the introduction of dialog curiosities closely contextualized with certain flows of a conversation, to improve user satisfaction/engagement.

In this context, our contributions are twofold: first, we propose a manually curated dataset of curiosities for the recipes and DIY domains; second, we propose a robust method to naturally insert curiosities in dialogues[1]. An A/B test with over 1000 conversations, conducted with real Alexa users, showed that the proposed approach achieves a relative rating improvement of 9.7%.

2 Curiosities Dataset

In this section, we explain the curiosities dataset creation process, that seeks to fill the existing gap

---

[1] `https://github.com/Mr-Vicente/Curiosity-Dataset`

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 149–155
September 11–15, 2023. ©2023 Association for Computational Linguistics

with regards to dialog curiosities for task assistants in the recipes and DIY domains. In particular, we considered the following principles: i) the curiosities' length matters significantly; ii) curiosities should be simple since dense and complex facts could have a negative impact on user engagement; and iii) the quality of each curiosity is more important than the number of curiosities.

2.1 Dataset Categories and Statistics

The dataset consists of a total of 1351 curiosities, with 754 curiosities for the cooking domain and 597 for the DIY domain, which are the target domains of the Alexa TaskBot challenge (Gottardi et al., 2022). Some examples of the curiosities general classes are listed below.

Sample Recipe concepts. Fruit (*e.g.* Avocado, Vitamin C); Meat (*e.g.* beef); Seafood (*e.g.* shrimp); tools (*e.g.* spatula); cuisine concepts (*e.g.* temperature); Popular countries' food (*e.g.* pizza, sushi); U.S. National food days.

Sample DIY concepts. American DIY statistics; DIY tools (*e.g.* hammer); Gardening (*e.g.* lawn mower); Garage (*e.g.* car, bike). House furniture (*e.g.* bookshelf); DIY tasks U.S. National days.

2.2 Curiosities Dataset Creation

The dataset was created by a manual process of searching and curating information found online. We started by considering a main class of a concept, for example "Fruit", and used Google search to find curiosities. After this first process, we get into more specific concepts, such as "Avocado". We complement our dataset with diverse temporally contextualized curiosities. Specifically, we employed a template-based approach to generate curiosities from national food days.

All the curiosities were manually curated to fit the characteristics and specifications identified, ensuring their quality and appropriateness for dialog and its domain.

2.3 Length per Curiosity

The length of a sentence can significantly affect the user's comprehension, especially in voice-based interactions, such as Alexa. Figure 1 shows the length distribution in words of the dataset for both domains. We deposited careful attention to conforming the curiosities length distribution to an average of 15 words, avoiding long sentences to maximize the readers' comprehension.

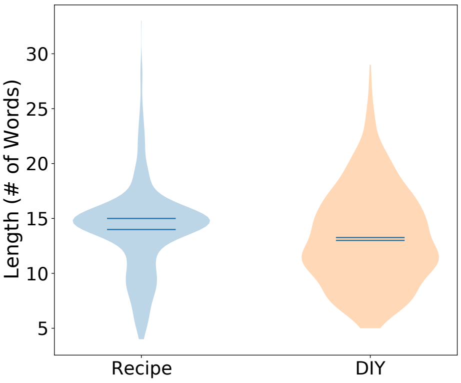

Figure 1: Curiosities length distribution.

3 Dialogues and Contextual Curiosities

One of the objectives of our work is to incorporate curiosities into a conversational assistant's flow while users are being guided through a task in natural language (Colas et al., 2020). However, inserting the right curiosity in the right dialog turn is a non-trivial task. Moreover, matching curiosities to a particular task with human effort would produce a high-quality reward, but, in practice, it is intractable due to the large amount of both tasks ($>$ 100k) and curiosities ($>$ 1k). To this end, we propose two distinct automatic approaches as described in the following sections.

3.1 Extracting Relevant Information

Recipes and DIY articles in their raw form consist of structured text composed of various elements such as the title, steps, and possibly extra information (*e.g.* ingredients, categories, and short descriptions). As a first step, we pre-process the tasks' content to match the curiosities by considering the most relevant content, taking into account the current phase of the conversation. Thus, for the recipes, we extracted the title, the steps, and the ingredients, whereas, for the DIY tasks, we extracted only the title and steps. The main goal is to capture fine-grained and task-specific details from each task, that will allow us to link a task to highly relevant curiosities, as described next.

3.2 Matching Curiosities to Dialogues

Given the information extracted from the tasks, we considered two approaches to match the curiosities to the conversation: (1) a text-based method, and (2) a semantic similarity search using pre-trained LM embeddings (Reimers and Gurevych, 2019).

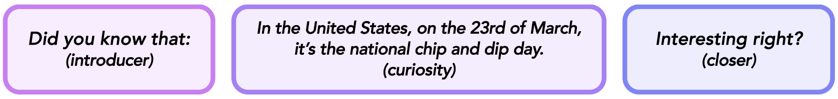
Did you know that:
(introducer)
In the United States, on the 23rd of March, it's the national chip and dip day.
(curiosity)
Interesting right?
(closer)

Figure 2: Example of a curiosity enclosed by an *opener* and a *closer*.

Text-based Method. In this approach, we first process the curiosities and the tasks' text by removing punctuation, stopwords, verbs, plurals, and domain-specific common words such as "hours" and "degrees". Then, we calculate two bag-of-words vectors considering the domain-specific words and both the curiosity and the current step of the task. After this, we perform the cosine similarity between both of these vectors for each curiosity available in the dataset and re-rank them according to this score.

Semantic Similarity Search Method. Matching a curiosity to the context of a dialogue requires some level of language understanding that goes beyond keyword matching. Hence, we considered a two-stage process considering a Sentence-BERT (Reimers and Gurevych, 2019) model that first separately encodes the task's content and all curiosities to map them to a common embedding space, allowing the assessment of the similarity between both. This is followed by a re-ranking cross-encoder method to further improve the results. Details of the algorithm are in Appendix B.

3.3 Inserting Curiosities in Dialogue

In a conversational task assistant, the primary objective is to assist the user in accomplishing a task (Gottardi et al., 2022). Therefore, the introduction of curiosities in the conversation should improve the dialog flow and maximize engagement. This requires a careful and contextualized blend of curiosities throughout the conversation.

Curiosity Offer/Backoff. Curiosities should improve the user experience, without negatively affecting the quality of a dialogue (Zheng et al., 2021). To ensure the overall users' satisfaction, and avoid non-intrusive behaviors, we designed a dialogue curiosity offer/backoff strategy (see Appendix C for the full algorithm).

An important aspect of our offer/backoff strategy is that we consider the user's cognitive load, and we never introduce curiosities at the beginning of a dialog, or when the user is listening to long steps ($\geq$200 words). This aims to keep the user focused, to provide short responses that account for the users' attention span. At these points of the dialogue, there are multiple voice instruction commands being explained to the user. Prompting and telling a curiosity would only cause confusion and cognitive overload.

We opted to ask the user at the end of a task step if they want to hear a curiosity (Appendix A, blue text). Given the question, the user can accept, deny, or ignore the request. If the user denies or ignores the curiosity, we opted to not prompt the user again, since the user might not have interest in this feature or may become frustrated. If the user accepts the curiosity, the bot responds with a fun fact following the structure discussed next.

Curiosities Openers and Closers. To smoothly insert individual curiosities in the dialog flow, while keeping the conversational gist, we propose a curiosity-to-dialog scheme, that encompasses curiosity linguistic *openers* and *closers*. To deliver a curiosity with the right tone of voice, we select an *opener* from a pre-defined list, to introduce the curiosity. Similarly, to gracefully end the insertion of a curiosity, we appended a *closer* phrase after the curiosity sentence. Given that the *closer* needs to act as a bridge between the curiosity and the main dialog flow, we formulated a set of ending sentences for the terminator phrase, with the aim of making them sound exciting, while signaling the end of the curiosity sub-flow. An example of a curiosity along with its corresponding *opener* and *closer* phrases, is illustrated in Figure 2.

4 Experimental Results

In this section, we detail the A/B testing setup and discuss the obtained results.

4.1 A/B Testing Setup

To measure the impact of introducing curiosities in a conversation, we performed A/B testing with Alexa device users, in the context of the Alexa Prize TaskBot Challenge 2021 (Gottardi et al., 2022).

The implemented dialogue system interacted with thousands of real users (Ferreira et al., 2022). The dialog state tracking is based on a BERT intent detector (Tavares et al., 2023) and the task retriever

Table 1: A/B testing results: system A engaged users in curiosities and system B had no possibility of curiosities at all. In system A, the user can accept, deny or ignore the curiosity recommendation.

Sys	User action	Conversations	Rating
A	Accepted (≥ 1)	526 (50.8%)	3.94
	Not-accepted	211 (20.4%)	3.55
B	Curios. disabled	299 (28.9%)	3.62

is based on a conversational search method (Ferreira et al., 2021). At the end of a conversation, the user is prompted to give a 1 to 5 rating regarding the quality of the conversation. We use the ratings as the success metric of the proposed work. We performed this study using an A/B testing method, by considering a version of the system with curiosities (A) and without curiosities (B). To ensure that we had high-quality data, we only considered conversations with a minimum of 3 turns, resulting in a total of 1036 conversations.

4.2 Dialogue Curiosities A/B tests

In Table 1, we summarize the A/B testing results that we conducted. We had 71.1% of the conversations in system A and 28.9% in system B. In system A, the user had the option to hear the curiosity and to decline it. Hence, 50.8% of the conversations had curiosities and 49.2% had no curiosities. In all systems, users were anonymous and randomly assigned to our system. Table 1 also relates the users' acceptance of curiosities to average ratings. The results show that users that accept at least one curiosity give on average a higher rating (3.94) compared to users that are not interested or that simply ignored the curiosity (3.55). Overall, this increase in rating shows that users that interact with the curiosities appear to be more engaged in the conversation, which in turn leads to a higher rating.

4.3 Ratings per Number of Curiosities

In this section, we examine system A results in more detail. Overall, we observed a positive result with 70% accepting a curiosity, 18% ignoring (the user does not confirm, *e.g*. "next step"), and 12% denying. Moreover, the relation between the number of curiosities per conversation and the rating is another positive result, Table 2. From these results, we can see that when curiosities are present in a dialogue, the rating is consistently higher than when no curiosities are said. In particular, we see a rating improvement from 3.55 against 3.74 in the worst-case scenario, and 4.13 in the best scenario. The mode is one curiosity per conversation, which corresponds to an average rating of 3.95, i.e. a relative improvement of 9.7%. These are encouraging results, showing that the users are receptive to listening to curiosities in the conversation which in turn leads to increased user satisfaction.

Table 2: System A's results breakdown: the number of provided curiosities and average rating.

Curiosities	None	1	2	≥ 3
Conversations	211	479	32	15
Avg. Rating	3.55	3.95(+9.7%)	3.74	4.13

4.4 Ratings by Curiosities Matching Method

We also examined the impact of the dialogue-curiosity matching methods of Section 3.2. Table 3 shows the results obtained with both methods. The two methods achieve high ratings, with the Semantic Similarity method obtaining slightly higher ratings, thus being more preferable.

Table 3: Rating by curiosity matching method.

Method	Count	Rating
Semantic Similarity	344 (64.18%)	3.99
Text-based	192 (35.82%)	3.86

5 Critical Discussion and Limitations

Manually Curated vs Hallucinated Curiosities. As an alternative to manually curated curiosities, current LLMs can generate curiosities contextualized to the conversation. We tested this strategy but observed that, often these curiosities are false and incorrect. Hence, this is not a viable solution when the dialog system guides a user through a complex manual task where reliability is key. An example of a false hallucination that we observed is "*Microwaves don't heat the food, they heat the water molecules in it, this causes them to vibrate which is what causes the heat. This vibration is good for your body because it causes your cells to produce more energy.*".

Long-term Effect of Curiosities. We studied the effect of curiosities during a period of 6 months with a controlled A/B testing. However, due to privacy issues, we did not track users, preventing us from studying the long-term effects of curiosities in recurring users.

Selection bias. Our study is limited to users that own an Alexa device, and to users that participated in the Alexa Prize TaskBot challenge. Moreover, in this setup, we only give fun facts to users who accept the offer of a fun fact. This creates a slight "selection bias" because users who accept a fun fact were probably the ones who were already enjoying the interaction and might have been more likely to give a higher rating. Likewise, users who rejected the curiosity were probably not enjoying the interaction and may have lowered their rating.

To obtain a reference rating (a neutral baseline), we disabled the fun facts functionality and tested the system. Table 1 provides an analysis that sheds some light on this issue. With fun facts disabled, the average rating is 3.62; users who refuse the fun fact, rate the system -0.07 points lower than the neutral baseline; users who accept the fun fact generally rate the system +0.32 points above the neutral baseline. Given the setup, in the future, we will study ways of mitigating possible sources of selection bias.

6 Conclusions

In this paper, we presented a novel approach to introducing curiosities in conversations. Specifically, we curated a dataset of curiosities in the recipes and DIY domains and evaluated the impact of introducing these curiosities in real human-agent conversations in the Alexa TaskBot challenge. We assessed the impact of curiosities in a conversational task assistant setting, and the results allow us to conclude that introducing curiosities in a non-intrusive manner and in the context of the dialog can increase user engagement and improve their appreciation of the dialogue system.

These findings have important implications for the design of conversational systems and can inform future research on incorporating curiosities in conversations to enhance the user experience. As future work, we will (1) investigate the use of generative models for creating factually grounded curiosities (Ouyang et al., 2022; Touvron et al., 2023) and compare them to manually curated curiosities; and (2) investigate methods that can contextualize the curiosities according to a graph of entities (Gonçalves et al., 2023).

Acknowledgments

This work has been partially funded by the FCT project NOVA LINCS Ref. UIDP/04516/2020, by the Amazon Science - TaskBot Prize Challenge and the CMU|Portugal projects iFetch CMUP LISBOA-01-0247-FEDER-045920), and by the FCT Ph.D. scholarship grant UI/BD/151261/2021. Any opinions, findings, and conclusions in this paper are the authors' and do not necessarily reflect those of the sponsors.

References

D. E. Berlyne. 1966. Curiosity and Exploration. *Science* 153, 3731 (1966), 25–33. `https://doi.org/10.1126/science.153.3731.25`

Cédric Colas, Tristan Karch, Nicolas Lair, Jean-Michel Dussoux, Clément Moulin-Frier, Peter Dominey, and Pierre-Yves Oudeyer. 2020. Language as a cognitive tool to imagine goals in curiosity driven exploration. *Advances in Neural Information Processing Systems* 33 (2020), 3761–3774.

Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. 2020. See, hear, explore: Curiosity via audio-visual association. *Advances in Neural Information Processing Systems* 33 (2020), 14961–14972.

Rafael Ferreira, Mariana Leite, David Semedo, and João Magalhães. 2021. Open-Domain Conversational Search Assistant with Transformers. In *ECIR (1) (Lecture Notes in Computer Science, Vol. 12656)*. Springer, 130–145.

Rafael Ferreira, Diogo Silva, Diogo Tavares, Frederico Vicente, Mariana Bonito, Gustavo Goncalves, Rui Margarido, Paula Figueiredo, Helder Rodrigues, David Semedo, and Joao Magalhaes. 2022. TWIZ: A conversational Task Wizard with multimodal curiosity-exploration. In *Alexa Prize TaskBot Challenge Proceedings*.

Gustavo Gonçalves, Joao Magalhaes, and Jamie Callan. 2023. Conversational search with random walks over entity graphs. In *ACM ICTIR*. ACM.

Anna Gottardi, Osman Ipek, Giuseppe Castellucci, Shui Hu, Lavina Vaz, Yao Lu, Anju Khatri, Anjali Chadha, Desheng Zhang, Sattvik Sahai, Prerna Dwivedi, Hangjie Shi, Lucy Hu, Andy Huang, Luke Dai, Bofei Yang, Varun Somani, Pankaj Rajan, Ron Rezac, Michael Johnston, Savanna Stiff, Leslie Ball, David Carmel, Yang Liu, Dilek Hakkani-Tur, Oleg Rokhlenko, Kate Bland, Eugene Agichtein, Reza Ghanadan, and Yoelle Maarek. 2022. Alexa, let's work together: Introducing the first Alexa Prize TaskBot Challenge on conversational task assistance. *Alexa Prize TaskBot Challenge Proceedings*.

C. Kidd and B. Y. Hayden. 2015. The Psychology and Neuroscience of Curiosity. *Neuron* 88, 3 (Nov 2015), 449–460.

Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondřej Kobza, Lenka Hýlová, and Jan Šedivý. 2021. Alquist 4.0: Towards Social Intelligence Using Generative Models and Dialogue Personalization. arXiv:2109.07968 [cs.CL]

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR* abs/2203.02155 (2022). `https://doi.org/10.48550/arXiv.2203.02155` arXiv:2203.02155

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. `https://arxiv.org/abs/1908.10084`

Diogo Tavares, Pedro Azevedo, David Semedo, Ricardo Sousa, and Joao Magalhaes. 2023. Task Conditioned BERT for Joint Intent Detection and Slot-filling. In *Progress in Artificial Intelligence - 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial, Portugal, September 5 - 8, 2023, Proceedings*. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023). `https://doi.org/10.48550/arXiv.2302.13971` arXiv:2302.13971

Qiong Wu, Chunyan Miao, and Zhiqi Shen. 2012. A curious learning companion in Virtual Learning Environment. In *FUZZ-IEEE 2012, IEEE International Conference on Fuzzy Systems, Brisbane, Australia, June 10-15, 2012, Proceedings*. IEEE, 1–8. `https://doi.org/10.1109/FUZZ-IEEE.2012.6251362`

Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao, and Chongjie Zhang. 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems* 34 (2021), 3757–3769.

A Example Dialogues

Table 4: Example of dialogue with curiosities. We highlight the question, opener, and closer.

System	**User**
Hi, this is an Alexa Prize Taskbot. I'm excited to help you with recipes or DIY tasks! Try asking me: How to cook salmon, or How to Decorate a Pot.	I want to make chicken tacos.
Sure, I found Chicken Tacos recipes in the Whole Foods Market. Would you like to 'start', or 'hear the next recipes'.	Start the recipe.
Chicken Tacos will be yummy. Step 1: Preheat oven to 375°F. To keep going with this recipe, say 'next'.	Next.
Step 2. Brush butter onto both sides of the tortillas. And look, would you like me to astonish you with a fun fact?	Sure.
Did you know that: The word tortilla comes from the Spanish word torta which means round cake. I love sharing these facts!	Next Step.

Table 5: Curiosity in the context of the recipe "Buttermilk Biscuits".

Step Text and Question
Step 2. Fold dry ingredients into buttermilk with your hands until a sticky dough forms. And look, would you like me to astonish you with a fun fact?
Fun Fact
How crazy is it that: In the US, on the 14th of May, it's the national buttermilk biscuit day. I found this really interesting! Anyway, to continue, say 'next'.

Table 6: Curiosity in the context of the DIY task "How to Braid Hair".

Step Text and Question
Step 3. Start with a secure base. If you tie hair into a ponytail with a hair tie, your braid will be easier to handle. And by the way, would you be interested in a fun fact about this?
Fun Fact
Alert! Alert! Fun fact time! The average person has between 100k and 150k strands of hair. This blew my mind! Anyway, to continue, say 'next'.

B Semantic Similarity Curiosity-Matching Algorithm

Algorithm 1: Curiosity Matching

```
Input  : $Tasks$: List of tasks
Input  : $n \leftarrow 10$: int (top-$n$ candidate curiosities)
Input  : $m \leftarrow 3$: int (top-$m$ candidate curiosities-task matches)
for each task in Tasks do
    Separate task's content into title, steps (and ingredients) using special tokens;
    Encode the task's content;
    Encode the domain-specific curiosities;
    Calculate the cosine similarity between the task's content and the curiosities;
    Select the top-$n$ curiosities;
    Apply a Cross-Encoder model to all $n$ pairs and select the top-$m$ pairs;
end
```

C Curiosities Offer/Backoff Algorithm

Algorithm 2: Curiosities Offer/Backoff

```
Input  : $T$: Task
Input  : $n\_steps$: int
Input  : $curr\_step$: int
Input  : $last\_fact\_step$: int
Input  : $questions\_asked$: int
Output : $ask\_curiosity$: bool
$k \leftarrow 6$;
$max\_questions \leftarrow (n\_steps \;//\; k) + 1$;
if $questions\_asked \geq max\_questions$ then
    $ask\_curiosity \leftarrow$ False;
else if $curr\_step \neq 1$ and
  $curr\_step = last\_fact\_step + k$ and
  $last\_fact\_step \leq curr\_step$ and
  $curr\_step \neq (n\_steps - 1)$ then
    $ask\_curiosity \leftarrow$ True;
else
    $ask\_curiosity \leftarrow$ False;
return $ask\_curiosity$;
```