A ADDITIONAL ILLUSTRATIVE EXAMPLES

As supplementary illustrations of the processes described in the main paper, we show three figures in this section: 1) illustration of the process of search related branch of inference in Figure 2, 2) a graphical model of all branches of the intermediate steps in Figure 3, and 3) a conceptual illustration of curricullum learning described in section 6 in Figure 4.

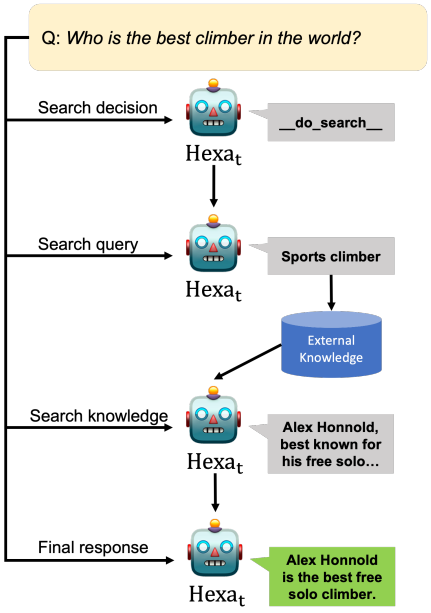

Figure 2: Example of external knowledge-grounded inference of our model. Here, we show an illustrative example of how the model inferences intermediate steps for external knowledge-grounded dialogue response generation. Following the same scheme as BB3 (Shuster et al., 2022b), given an input context, with a special token *__is-search-required__*, the model decides whether to search or not by outputting *__do-search__* or *__do-not-search__*. Upon deciding to search, the model then generates a search query that will be used in the external knowledge source such as web, to retrieve relevant documents. For the query generation, a special token of *__generate-query__* is appended at the end of the original context. With the retrieved documents, the model then generates a knowledge piece for the context using a special token *__generate-knowledge__*. Finally, with the generated knowledge appended to the context, the model generates the response for the given context.

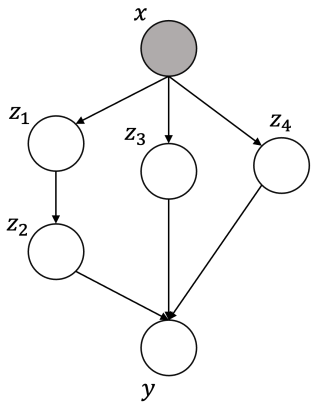

Figure 3: Graphical model of latent variables. Given the dialogue context x, $z_1 \sim p(\cdot|x;\theta)$ and $z_2 \sim p(\cdot|x, z_1;\theta)$ are the search query and the search knowledge respectively, where the search query is used as a query to retrieve external knowledge from sources such as web and the search knowledge is generated based on the retrieved external knowledge and x. $z_3 \sim p(\cdot|x;\theta)$ is the entity knowledge, generated using only the dialogue context x. Finally, $z_4 \sim p(\cdot|x;\theta)$ is the retrieved dialogue history-based internal knowledge, conditioned on x. After generating these intermediate steps, the final response $y \sim p(\cdot|x, z_{2:4};\theta)$ is conditionally generated.

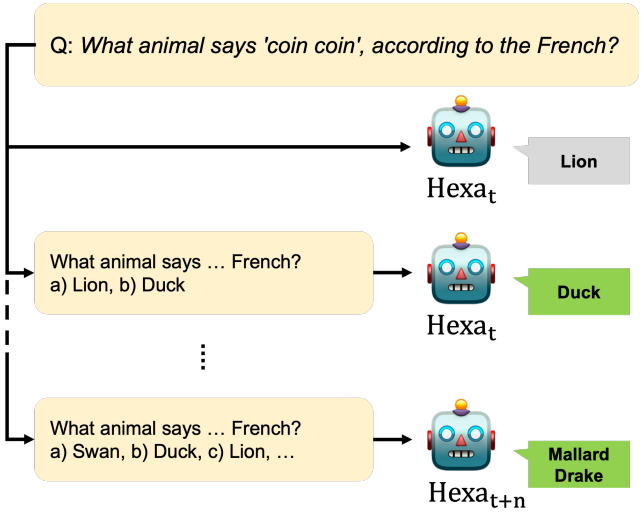

Figure 4: Conceptual illustration of curriculum learning in Hexa. Here, a question of *What animal says 'coin coin', according to the French?* with the ground truth *Duck* is given. The model at t produces a wrong response *Lion* but attempts again with a guided prompt $h_i^t = \{\text{Duck}, \text{Lion}\}$, and the response is correct. After n iterations, the model is asked with the same question again with a expanded set $h_i^{t+n} = \{\text{Swan}, \text{Duck}, \text{Lion}, ...\}$ and outputs *Mallard Drake*. Since *Mallard Drake* is a species of Duck, it can also be one of the ground truth output, and Hexa includes it in the training set.

B MODULE-WISE EVALUATION

As described in subsection 5.3 of the main paper, we report the module-wise evaluation for the instances where different models share the same decision paths. Note that memory-related modules used in multi-turn conversation scenarios were excluded from this experiment since it is impossible to compare the results under same condition, i.e., using exactly same memory of the conversation. The results in Table 7 show Hexa achieving the highest scores in search and entity knowledge generation. Combining this result with that of Table 1, we may draw an hypothesis that the performance in search and entity knowledge generation has relatively higher correlation with the performance of final response generation compared to the other two.

Table 7: Module-wise evaluation.

Model	Search Query R-L	Search Knowledge R-L	Entity Knowledge R-L	Search Decision Accuracy
BB3	51.76	25.11	13.26	76.84
BB3-SL	**52.40**	21.90	20.07	**79.52**
STaR	48.88	24.81	22.17	77.76
Hexa	46.41	**26.18**	**24.35**	77.99

C ANALYSIS ON BOOTSTRAP

C.1 BOOTSTRAP SAMPLES

As mentioned in section 4 of the main paper, composing the guided prompt only with the ground truth, as in STaR, may collapse to simply copying the response throughout the intermediate steps, which can degenerate the generalization ability of the model. In order to empirically present such phenomenon, we compare the generated samples of search query and knowledge generation modules among different methods. We specifically analyze the search query samples from TriviaQA (Joshi et al., 2017) as this task is knowledge intensive QA with short responses where the copying phenomenon should be more easily observable. We report the rate of the number of search queries that include copies of the ground truth in bootstrapped samples with the guided prompt. As shown in the row labeled with *Search query copy rate* of Table 8, STaR's copy rate of search query is approximately twice the value of Hexa. Similarly, we also report the average overlap score by ROUGE-L between the generated knowledge and the ground truth in the bootstrap samples on all tasks except TriviaQA. As shown in the row labeled with *Knowledge overlap score* of Table 8, Hexa generates knowledge more dissimilar to the ground truth compared to STaR.

Table 8: The *Search query copy rate* (%) and the *Knowledge overlap score*. The average values across all iterations are presented.

	STaR	STaR w/o hint	Hexa
Search query copy rate	18.42	5.37	9.24
Knowledge overlap score	15.33	10.94	11.47

We observe that a variant of STaR that does not use the guided prompt, labeled with *STaR w/o hint* in Table 8 is inferior to Hexa even though it has lower values on both the copy rate and the overlap score than Hexa (see Table 5 of the main paper). This implies that reducing the copy rate or the overlap score may not be the direct cause of the improvement and the falsely predicted responses in the guided prompt of Hexa may make the bootstrap better than STaR. We hypothesize that the guided prompt of Hexa can provide reusable knowledge for the model to generate more informative response to the given context.

Although a generated response may not pass the matching function threshold, parts of it can contain information relevant to answering the given question. When this falsely predicted response is augmented in the guided prompt along with the ground truth, the guided prompt as a whole contains more relevant knowledge than just the ground truth, leading to generating more informative responses.

An example of such case is presented in Table 9 where the information from the falsely predicted response is combined with the ground truth in the final response.

Table 9: A bootstrap sample of Hexa on FITS Xu et al. (2022c). In this case, the first response is classified to the falsely predicted response (**Middle**). Then, Hexa tries to produce the response again along with the guided prompt that contains ground truth and the 1st response. In this 2nd try, the model generates the copy of ground truth into the knowledge. Along with the knowledge, the model generates the response which is classified to correct response and bootstrapped. The 2nd response also contains some part of the 1st response highlighted in olive (**Bottom**).

Input	I'm trying to learn about sea turtles, could you help me find a list of places around the world where they live?
Ground truth	Green sea turtles can be found in more than 140 countries worldwide while the nesting grounds are found in 80 countries.
1st response (Falsely predicted)	Sea turtles are found all around the globe. Among the most common places they· are listed are Belize, where they can be found on a regular basis, and the southern U.S./Southwest Mexico coast. They can still be seen in the Gulf as well, including florida, alaska and flores islands
Query	List of sea turtle habitat
Knowledge	Green sea turtles are found in more 140 countries worldwide while the nesting grounds are located in 80 countries
2nd response (Bootstrapped)	Green-sea turtle can be seen near the gulf of mexico and you can still find blue · turtles in the southeast. They have found them in 80 countries, including florida, alaska and flores islands

C.2 Bootstrapping Rate

The bootstrapping rate, the number of bootstrapped data divided by the number of attempted instances, will be different for models depending on the used guided prompts. For example, STaR w/o hint, a version of STaR that does not take any guided prompt, may have a lower bootstrapping rate since it could be difficult to generate response similar to the ground truth without the guidance. To verify, we obtain the average bootstrapping rate across the iterations for different models with different guided prompts and the results are shown in Table 10. Interestingly, we find that Hexa has the highest bootstrapping rate, which greatly enhances the bootstrap data collection speed. More interestingly, Hexa w/o ground truth which only uses the unmatched responses for the guidance has better bootstrapping rate than STaR which only uses the ground truth for the guidance. This suggests that the self-generated responses are indeed meaningful information that correctly guides the response generation.

Table 10: Comparison for bootstrapping rate (%).

	STaR	STaR w/o hint	Hexa	Hexa w/ random hint	Hexa w/o ground truth
Bootstrapping rate	22.02	7.61	**29.82**	22.79	23.58

D Evaluation on Totally-Unseen Tasks

Here, we provide an additional evaluation to test robustness of the methods. We consider OpenDialKG (Moon et al., 2019), a conversational reasoning benchmark dataset, consisting open-ended conversations between humans. In this task, the system is demanded to recommend items that users might prefer through multi-turn conversations on various domains including movies, books, sports, and music. Note that this task is not included during both BB3-training and finetuning (e.g. Hexa). As shown in Table 11, Hexa outperforms the other baselines in automatic evaluation for this totally-unseen task as well.

Table 11: Results for OpenDialKG unseen during BB3-training and finetuning.

Model	F1	R-L
BB3-60K	15.02	13.84
BB3-SL	15.46	14.41
STaR	15.68	14.76
Hexa	**18.08**	**16.60**

E DIVERSITY OF RESPONSES

Here, we conduct an automatic evaluation for diversity between final responses. There exists a tradeoff between the diversity and correctness, as group of correct answers would tend to resemble each other compared to set of random answers. Therefore we specially design a method to measure the appropriate diversity within a certain boundary of correctness. We first randomly sample intermediate steps z and y 10 times for each instances. Then, we select samples that satisfy the matching function. Furthermore, we compute *Self-BLEU* (Zhu et al., 2018) and *Distinct* (Li et al., 2016) scores for the set of selected samples. Table 12 shows *Matching rate* of the samples, Self-BLEU (quadrigram), and Distinct (bigram) scores of the matching samples for BB3-SL, STaR, and Hexa on seen KGD tasks. The result shows that Hexa produces most matching answers and achieves better performance in terms of diversity, indicating the capability of producing more diverse correct responses.

Table 12: Comparison of correctness and diversity of final responses between finetuning methods.

	Matching rate ↑	Self-BLEU ↓	Distinct ↑
BB3-SL	11.98	92.16	11.19
STaR	12.92	92.75	10.81
Hexa	**13.98**	**91.88**	**11.51**

F FURTHER DISCUSSION AND LIMITATIONS

Bootstrap Quality An overlaying assumption in the self-improving methods such as Hexa is that samples with irrelevant z would not be bootstrapped since they are unlikely to lead to appropriate responses. However, in practice, those cases may be included in the bootstrap and deteriorate the self-learning process. The current design does not include a mechanism to prevent this issue but a straight forward solution to such problem is to include a rejection sampling. For example, upon sampling an intermediate step z, we can decide to reject the sample if its presence and absence does not change the final response, meaning it has no relevance in producing the final response. This method can easily be extended to Hexa and we leave it as a possible candidate of future work direction.

G ADDITIONAL STUDY

G.1 THRESHOLD SELECTION

Before training, as described in subsection 5.2 of the main paper, we conduct a *task-wise threshold selection* that greedily searches the threshold value on each task to choose the appropriate threshold value, and use the selected threshold values (see Table 13) during training. We expect that this task-specific selection can lead to the performance improvement since undesired target responses can be bootstrapped when we inappropriately use a low threshold for the task, and only a narrow set of responses can be bootstrapped when we inappropriately use a high threshold for the task. To show these, we compare the performance between the *task-wise threshold* and *fixed threshold* $\in \{0.1, 0.25, 0.3, 0.4\}$ that uses the same threshold value for all tasks except TriviaQA (Joshi et al., 2017) which used the threshold value of 0.99 as in the task-wise threshold. As shown in Table 14, the use of low threshold $b = 0.1$ degrades the overall performance while the use of high threshold $b = 0.4$ degrades the performance on unseen tasks. The use of median value $b = 0.25$ or approximate average value $b = 0.3 \approx 0.2929$ is inferior to the use of the task-wise threshold.

Table 13: Different threshold values used for each task

Task	Threshold
Question Answering	
TriviaQA Joshi et al. (2017)	0.99
MS Marco Nguyen et al. (2016)	0.25
Knowledge-Grounded Dialogue	
WoW Dinan et al. (2019)	0.25
WoI Komeili et al. (2022)	0.25
FITS Xu et al. (2022c)	0.35
Open-Domain Dialogue	
PersonaChat Zhang et al. (2018)	0.35
Multi-Session Chat Xu et al. (2022a)	0.25
Task-Oriented Dialogue	
GoogleSGD Lee et al. (2022)	0.35

Table 14: Comparison between the *task-wise threshold* and *fixed threshold*.

	Seen		Unseen	
Threshold	F1	R-L	F1	R-L
Task-wise	**20.83**	**22.25**	**18.55**	**16.63**
Fixed, $b = 0.1$	18.69	21.43	16.88	15.28
Fixed, $b = 0.25$	20.46	21.8	18.01	16.12
Fixed, $b = 0.3$	**20.8**	22.07	18.16	16.24
Fixed, $b = 0.4$	20.63	**22.3**	16.88	15.4

G.2 MATCHING FUNCTION

As discussed in the main paper, the matching function B could be essential in the proposed method. Here, we consider an alternative choice of similarity function called *Sentence-BERT* (S-BERT) (Reimers & Gurevych, 2019). Sentence-BERT was trained to measure the semantic similarity between two sentences and therefore can be used distinguish correct answers according to semantic similarity rather than the overlap. We replace ROUGE-L in matching function of Hexa with the cosine-similarity score between the S-BERT embeddings of the ground truth and the generated response. We label this setting as *Hexa w/ Sentence-BERT* in Table 15.

The results in Table 15 show that even with Sentence-BERT, Hexa achieves competitive scores in all metrics in both seen and unseen tasks, all higher than that of BB3-SL. Furthermore, we can observe that Hexa with ROUGE-L even improves in S-BERT score. Upon this observation, we conclude that ROUGE-L is effective and efficient choice of matching function, as S-BERT requires additional model inference to calculate the score.

Table 15: Ablation on similarity function. The Sentence-BERT score is denoted as S-BERT. Superscripted by * as the default setting for Hexa. The result of BB3 is added for reference.

	Seen			Unseen		
Model	F1	R-L	S-BERT	F1	R-L	S-BERT
BB3	17.83	19.48	41.37	15.91	14.29	34.37
BB3-SL	18.87	20.03	43.94	15.4	13.78	34.32
Hexa w/ ROUGE-L*	**20.83**	**22.25**	46.03	**18.55**	**16.63**	**36.72**
Hexa w/ Sentence-BERT	19.28	20.95	**46.56**	16.06	14.8	35.81

H EVALUATION WITHOUT MODULES

Here, in order to examine the knowledge stored in the pre-trained model BB3 and trained model by Hexa, we evaluate the system to generate the final response directly, bypassing the intermediate steps before the final response generation. The average scores across the tasks categorized as Seen of the main paper are shown in Table 16 and we see that skipping the intermediate steps underperforms BB3 and Hexa with the intermediate steps.

Table 16: Evaluation models with (denoted as *w/*) and without (denoted as *w/o*) modules.

Model	F1	R-L
BB3 w/o modules	17.05	17.26
BB3 w/ modules	17.83	19.48
Hexa w/o modules	19.14	19.25
Hexa w/ modules	20.83	22.25

I HUMAN EVALUATION

To gauge performance across multiple aspects of quality of responses of dialogue generation by Hexa and STaR, we conduct the human evaluation on the nine tasks of KGD, ODD, and TOD. We randomly select 180 samples (20 samples per task), and each sample is evaluated by ten different human experts. Two responses with the generated knowledge from each model are shown to the annotators for each instance. For A/B testing on Fluency, Relevance, and Faithfulness, we give one score to the model if it's response is received an equally good or better than the other one, otherwise, we give zero score to the model. Figure 5 display the annotator instructions for the three aspects.

J CODE SUBMISSION

Our code for Hexa is involved in the supplementary meterials. Due to the file size limit, we include only the training and evaluation code that runs on a small subset of the original training and test data. We will release our entire code that runs on the full training and test data, and the trained model file upon acceptance.

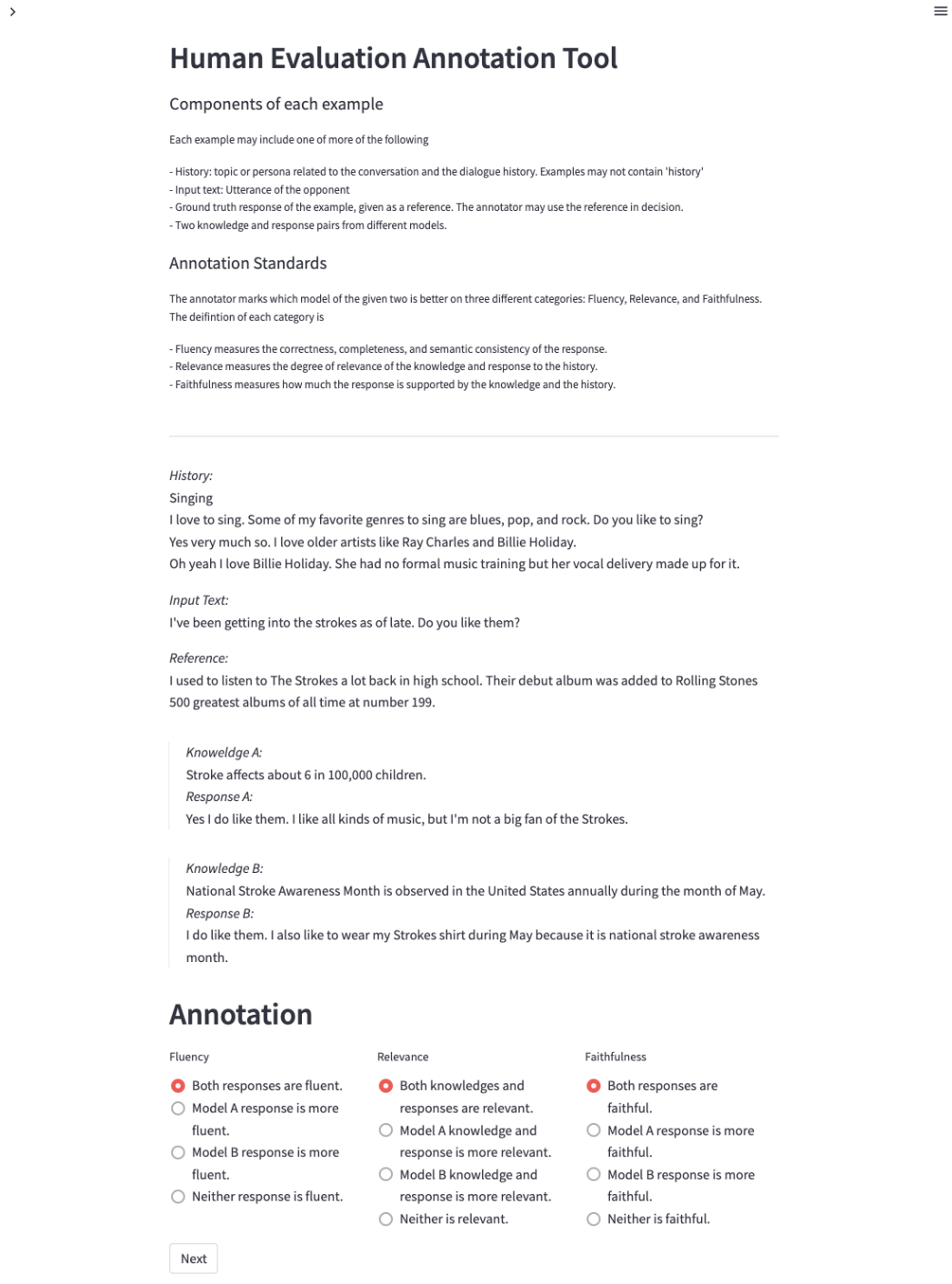

Figure 5: Screenshot of one instance in the annotation tool used for human evaluation.

K GENERATED SAMPLES

Table 17: Generated examples from WoW Dinan et al. (2019).

Input	What is the most visited paris landmark?
query	Most visited place in paris
knowledge	Catacombs of Paris
BB3-SL	The catacombs are the most popular landmark in Paris.
query	Most visited place in paris
knowledge	Eiffel Tower
STaR	Eiffel tower.
query	Most visited place in paris
knowledge	The Eiffel Tower
Hexa	The eiffel tower is a landmark in paris.
Input	[...] Yes, can you find out more about the very first ncaa football game, in 1869? The first ncaa football game was played on nov 6, 1869, between rutgers and princeton. Rutgers won. Great. One last question: what was the score of that game?
query	Rutgers v princeton ncaafb game score
knowledge	Rutgers won the game by a score of 6 runs to princeton's 4.
BB3-SL	Rutgers won that game by 6 runs.'
query	Rutgers princeton game score
knowledge	Rutgers won the game by a score of 6 runs to princeton's 4.
STaR	Rutgers won that game by 6 runs.
query	Rutgers princeton nca football game 1869
knowledge	Rutgers won the game by a score of 6 to 4.
Hexa	Rutgers won their first game against princeton by a 6-4 score.
Input	What do they think jupiter's rocky core is made of?
query	Jupiter rocky core composition
knowledge	Gas giants consist mostly of hydrogen and helium.
BB3-SL	Jupiter is a gas giant, and gas giants consist of mostly hydrogen.
query	Jupiter rocky core
knowledge	Metallic hydrogen
STaR	Metallic hydrogen
query	Jupiter composition
knowledge	Jupiter's rocky core is made of liquid metallic hydrogen.
Hexa	Jupiter's core is believed to be made of metallic hydrogen

Table 18: Generated examples from WoI Komeili et al. (2022).

Input	Usually for 10-12 minutes, I like to get a little color on them. What is your favorite meal to make?
Input Persona	your persona: I like pasta. I know how to cook sausage. partner's persona: I enjoy cooking. I have a freezer. I like breakfast burrito.
query	pasta
knowledge	Banana pasta Banana pasta is a type of pasta prepared using banana as a primary ingredient. The product is typically made with unripe, [...]
BB3-SL	My favorite pasta is banana pasta. It is a pasta made with banana.
query	pasta
knowledge	Banana pasta is cooked like standard pasta, typically by boiling it.
STaR	My favorite pasta is banana pasta. It is cooked just like regular pasta.
query	pasta
knowledge	My favorite meal to cook is pasta.
Hexa	My favorite meal is pasta and I love to make it for my family.
Input	[...] Good for you! Are you watching the Superbowl this year? I will! It is on the 7th right ? It is. Are you a Brady fan or foe?
query	tom brady
knowledge	Tom Brady–Peyton Manning rivalry The Tom Brady–Peyton Manning rivalry is a series of games that took place between 2001 and 2016, involving two quarterbacks in the National Football League (NFL): [...]
BB3-SL	I am not a fan of Tom Brady, but I am a fan of Peyton Manning!'
query	tom brady
knowledge	The Tom Brady–Peyton Manning rivalry is a series of games that took place between 2001 and 2016, involving two quarterbacks in the National Football League (NFL): [...]
STaR	I am not a fan of Tom Brady but I am a fan of Peyton Manning
query	tom brady
knowledge	I am a fan of Tom Brady.
Hexa	I'm a big fan of tom brady! I like him a lot. I think he is a great quarterback.

Table 19: Generated examples from MSC Xu et al. (2022a).

Input	[...] I just wanted something different at the moment! Would you ever consider dying your hair? I don't think so, I'm way too old for that. I don't have much hair anymore! Well, I think you could still do something cool perhaps! Any ideas?
Input Persona	your persona: I volunteer at a soup kitchen. I am depressed. I grew up in poverty. I am old. partner's persona: I enjoy volunteer work. I like soup kitchens. I dyed my hair purple.
entity	something
BB3-SL	I'm not sure, maybe something with feathers?
entity	"
STaR	I have been thinking about getting a tattoo, but I'm not sure if I have the time or money for it.
entity	the holidays
Hexa	I would like to dye my hair pink for the holidays. I think it would be a good idea.

Input	[...] Good, I was finally able to find a job that valued it. Who knew economics wouldn't be easily employable. What would you go back to school for? Given your experience, possibly economics! Definitely something that doesn't involve hospitals. Haha I can only imagine how bad you need a break from hospitals. If you'd like I can show you how to coupon to help save money when you go back in school
Input Persona	your persona: I have a job. I am considering going back to school to get a higher paying job. I am in school. partner's persona: I have coupons. I won a coupon raffle. I have been to Disneyland. I have an economics degree.
entity	money
BB3-SL	That would be great! I'd love to learn how to save some money.
entity	economics
STaR	That would be great! I'd love to learn how to do that. I've never used coupons before. Do you have any tips for saving money in economics?
entity	school
Hexa	That would be great! I would love to learn how to do that. I would love to be able to save more money. I have been thinking about going to nursing school.

Table 20: Generated examples from Ms Marco Nguyen et al. (2016).

Input	why do electric dryers need to be vented
query	Electric dryer venting
knowledge	Hot dry(er) air enters the drum from either the front or the rear, and evaporates some of the moisture from the tumbling
BB3-SL	Hot dry air enters from the front and rear of the drum.
query	Electric dryer vent
knowledge	Hot dry(er) air enters the drum from either the front or the rear, and evaporates some of the moisture from the tumbling
STaR	Hot dry air enters from the front and rear of the dryer.
query	Electric dryer vent
knowledge	A clean, unobstructed dryer vent improves the safety and efficiency of the dryer.
Hexa	Dryer venting is important because it improves the efficiency and safety of dryers.
Input	amisom somalia
query	Buy amisom in somalia
knowledge	African Union Mission to Somalia (2007–present)
BB3-SL	African Union mission to Somalia
query	Amisom
knowledge	The African Union Mission to Somalia (AMISOM), guarding the palace, became involved in the fighting and was unable to take control of rebel strongholds.
STaR	The African Union mission to Somalia.
query	Somalia amisom
knowledge	The African Union Mission in Somalia (AMISOM) is an active, regional peacekeeping mission operated by the African Union with the approval of the
Hexa	The African Union mission in Somalia is a peacekeeping operation.
Input	why do puppies lick their owners
query	Puppy licking owner
knowledge	lick each other's faces in greeting, then they begin to sniff any moist membranes where odors are strongest i.e. mouth, nose, anal regions and urogenital areas.
BB3-SL	Puppies lick each other in greeting.
query	Puppy licking owner
knowledge	Licking can communicate information about dominance, intentions and state of mind
STaR	Licking is a way for puppies to communicate with their owners.
query	Puppy licking owner
knowledge	Licking can communicate information about dominance, intentions and state of mind, and like the yawn is mainly a pacifying behavior.
Hexa	Puppies lick their owner to communicate with them. Licking is a way for puppies to communicate with their owners.