Advancing Question Generation with Joint Narrative and Difficulty Control

Bernardo Leite and Henrique Lopes Cardoso
LIACC, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
{bernardo.leite, hlc}@fe.up.pt

Abstract

Question Generation (QG), the task of automatically generating questions from a source input, has seen significant progress in recent years. Difficulty-controllable QG (DCQG) enables control over the difficulty level of generated questions while considering the learner's ability. Additionally, narrative-controllable QG (NCQG) allows control over the narrative aspects embedded in the questions. However, research in QG lacks a focus on combining these two types of control, which is important for generating questions tailored to educational purposes. To address this gap, we propose a strategy for Joint Narrative and Difficulty Control, enabling *simultaneous* control over these two attributes in the generation of reading comprehension questions. Our evaluation provides preliminary evidence that this approach is feasible, though it is not effective across all instances. Our findings highlight the conditions under which the strategy performs well and discuss the trade-offs associated with its application.

1 Introduction

Question Generation (QG) focuses on the automated generation of coherent and meaningful questions targeting a data source, including unstructured text or knowledge bases (Rus et al., 2008). Controllable QG plays a crucial role in education (Kurdi et al., 2020), as it facilitates the generation of personalized questions that address the unique needs and learning goals of students. Recent work on QG utilized techniques such as fine-tuning (Zhang et al., 2021; Ushio et al., 2022) and few-shot prompting (Wang et al., 2022b; Chen et al., 2024) to generate questions based on a source text and, optionally, a target answer. In controllable QG, this process is augmented by incorporating controllability labels into the input or prompt to guide the generation process. Specifically, research on Narrative-Controlled Question Generation (NCQG) focuses on controlling the **content** of generated questions, guided by underlying narrative elements (e.g., causal relationship) (Zhao et al., 2022; Leite and Lopes Cardoso, 2023; Li and Zhang, 2024). In turn, Difficulty-Controllable Question Generation (DCQG) emphasizes controlling the expected difficulty in answering the questions (Gao et al., 2019; Kumar et al., 2019; Cheng et al., 2021; Bi et al., 2021). Some studies have considered the relationship between question **difficulty** and the **learner's ability** (Uto et al., 2023; Tomikawa and Uto, 2024).

Passage: Once there were a hare and a turtle. The hare was proud of his speed and challenged the turtle to a race. Although the turtle was slow, he accepted. The hare quickly left the turtle behind but decided to rest and fell asleep. Meanwhile, the turtle kept going steadily and eventually reached the finish line first, winning the race.

Narrative: "character" **Difficulty**: "easy"
Generated QA Pair: Who challenged the turtle to a race? The hare.

Narrative: "outcome" **Difficulty**: "medium"
Generated QA Pair: What happened after the hare left the turtle behind? Decided to rest and fell asleep.

Narrative: "outcome" **Difficulty**: "hard"
Generated QA Pair: What happened because the turtle kept going steadily? The turtle won the race.

Figure 1: Illustrative example of controlled question-answer generation with varying difficulty levels and narrative attributes.

However, research in controllable QG lacks the combination of these two types of control, which is especially important to facilitate human control (Wang et al., 2022a) in the ever-increasing usage of generative models in this field. Therefore, this research proposes a strategy that explores the feasibility of joining narrative and difficulty control to generate reading comprehension question-answer

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 647–659
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

(QA) pairs from children-targeted narrative stories. Figure 1 shows an example of the strategy. Formally, we investigate the following research question (RQ): *How effectively can we control the generation of question-answer pairs conditioned on both narrative and difficulty attributes using a modest*[1] *scale model?*

For our experiments, we use a well-known dataset — FairyTaleQA (Xu et al., 2022) — in which each question is already annotated with one of seven narrative labels. Our method involves two main steps: (1) using simulated-learner QA systems to answer questions from FairyTaleQA, thereby estimating the difficulty labels via Item Response Theory, and (2) applying a joint narrative and difficulty control model, utilizing human-annotated narrative labels and the estimated difficulty labels for each question.

The proposed method is evaluated to determine whether both NCQG and DCQG have been successfully applied to the generated questions. For NCQG, we compare the similarity between human-authored and generated questions. For DCQG, we assess the performance of simulated-learner QA systems on questions generated with distinct difficulty levels. Although the results demonstrate the effectiveness of the strategy, NCQG shows consistent success, whereas DCQG exhibits moderate success, with performance varying across specific narrative attributes and difficulty levels. Our goal is to highlight the conditions under which the strategy performs with high or low efficacy, providing insights for researchers pursuing similar research lines. In summary, our contributions are:

- We propose a joint strategy for controlling the generation of question-answer pairs conditioned on narrative and difficulty attributes.
- We report on the linguistic features influenced by control and conduct an error analysis of the generated QA pairs, providing insights into the performance and limitations of the method.

2 Background and Related Work

2.1 Controllable Question Generation (CQG)

As stated by Li and Zhang (2024), prior research on CQG has explored two main perspectives: content (or type) and difficulty.

Content control relates to the linguistic elements incorporated into the generated questions. For instance, Ghanem et al. (2022) proposed controlling specific reading comprehension skills, such as figurative language and vocabulary. Additionally, Zhao et al. (2022) focused on controlling narrative elements, while Leite and Lopes Cardoso (2023) extended this approach by controlling explicitness attributes. Elkins et al. (2023) propose to control Bloom's question taxonomy (Krathwohl, 2002).

Difficulty control is related to the challenge of answering the generated questions, a concept that is often subjective (i.e., difficulty can vary depending on the respondent). In this regard, Gao et al. (2019) assigned difficulty labels (easy or hard) to questions based on whether QA systems could answer them correctly and used these labels as inputs to control the generation process. Kumar et al. (2019) proposed estimating difficulty based on named entity popularity, while Bi et al. (2021) tackle the challenge of high diversity in QG. Furthermore, Cheng et al. (2021) controlled question difficulty by considering the number of inference steps required to arrive at an answer.

One limitation of previous approaches is (1) the lack of emphasis on the relationship between question difficulty and learner ability. Addressing this problem, Uto et al. (2023) proposed to use Item Response Theory (IRT) (Lord, 2012), a mathematical framework in test theory, to quantify question difficulty and directly relate it to learner ability. Another limitation is (2) the lack of integration of multiple attributes. While Li and Zhang (2024) combine both narrative and difficulty attributes, they define *difficulty* in terms of answer explicitness and the number of sentences needed to answer the questions. The novelty of this study lies in integrating content control, through narrative elements, with difficulty control *informed by simulated learners' ability*, thus building on the foundations laid by previous research.

2.2 Item Response Theory (IRT)

IRT (Lord, 2012) is a statistical framework used to study the interaction between test-takers (ability or proficiency) and their performance on test items. A key aspect of IRT is to model the relationship between question difficulty and learner ability, offering insights into how well a question differentiates between individuals with varying levels of skill. This relationship allows for an estimation of

[1]<1 billion of parameters.

the likelihood that a learner with a specific ability level can correctly answer a given question, making it particularly useful for adaptive testing and understanding question complexity. A commonly used model in IRT is the **Rasch model**, which assumes that the probability of a correct response depends on the relation between learner ability (θ) and the item's difficulty (b):

$$P(X_{ij} = 1 \mid \theta_i, b_j) = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}}, \quad (1)$$

where θ_i is the learner ability of individual i, b_j is the difficulty of item j, and $P(X_{ij} = 1 \mid \theta_i, b_j)$ is the probability that individual i correctly answers item j. In our study, we use IRT to estimate both question difficulty (b) and learner ability (θ) parameters.

2.3 FairyTaleQA: Purpose and Value

We use the FairytaleQA dataset (Xu et al., 2022) because its stories and corresponding question-answer pairs align with the goal of addressing *narrative comprehension*. According to Xu et al. (2022), narrative comprehension represents a high-level cognitive skill closely linked to overall reading proficiency (Lynch et al., 2008). A key feature of FairytaleQA is the expert annotations on each question, which are grounded in evidence-based frameworks (Paris and Paris, 2003; Alonzo et al., 2009). The annotated narrative elements targeted for control are:

- **Character**: Addresses identities or traits of story characters (e.g., "Who...?");
- **Setting**: Focusing on the time and place of events, often starting with "Where...?" or "When...";
- **Action**: Related to the actions of characters;
- **Feeling**: Exploring emotional states or reactions (e.g., "How did/does X feel?");
- **Causal relationship**: Addressing cause-and-effect (e.g., "Why...?" or "What caused/made X?");
- **Outcome resolution**: Focusing on the outcomes of events (e.g., "What happened/happens after X?");
- **Prediction**: Questions about future or unknown events based on textual evidence.

While there are other popular educational QA datasets (following the open-ended *wh*-questions format), such as NarrativeQA (Kočiský et al., 2018) and StoryQA (Zhao et al., 2023), they are not annotated with specific reading comprehension skills. This further motivated our decision to use FairytaleQA in this study.

3 Method

This section outlines the methodology of this research, which includes augmenting FairytaleQA with IRT-based difficulty labels and developing a question-answer pair generation model with joint narrative and difficulty control. Figure 2 provides an overview of the steps discussed in this section.

3.1 Augmenting FairytaleQA With IRT-Based Question Difficulty Labels

Let D be our dataset consisting of instances represented as quartets:

$$D_i = (t, q, a, n), \quad (2)$$

where t is a text, q is the question, a is the answer about the text, and n is the narrative element associated with the question-answer pair (q, a). The aim is to create a fifth element d, resulting in a new instance augmented:

$$D_{i-\text{augmented}} = (t, q, a, n, d), \quad (3)$$

where d is the estimated difficulty value associated with the question-answer pair (q, a). To create these augmented instances, we used the method proposed by Uto et al. (2023) and Tomikawa et al. (2024):

1. **Collecting response data for each question-answer pair**: We collected answers to the questions from multiple respondents. Given the unavailability of real students, we utilized simulated-learner QA systems, which are models capable of automatically extracting answers to the posed questions. As explained in Section 4.2, the QA models were deliberately chosen to represent different levels of performance to simulate varying ability levels.
2. **Estimating Question Difficulty with IRT**: Using the answers collected from the simulated-learner QA systems, we estimated the difficulty of each question using IRT, specifically employing the Rasch model as described in Section 2.2.

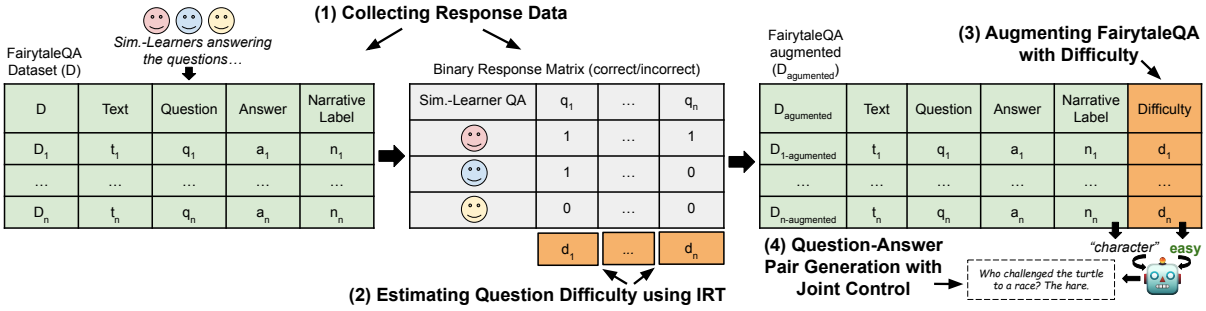

Figure 2: Overall methodology for joint narrative and difficulty control.

3. **Augmenting FairytaleQA with difficulty estimates**: Based on the estimated difficulty values, we augment each instance of the dataset with d, resulting in $D_{i-\text{augmented}} = (t, q, a, n, d)$.

3.2 Question-Answer Pair Generation with Joint Narrative and Difficulty Control

The controllable process can be represented as follows: given an instruction prompt p, the aim is to use a model M to generate a question-answer pair $(q_{\text{new}}, a_{\text{new}})$. This can be formulated as:

$$(q_{\text{new}}, a_{\text{new}}) = \text{M}(p), \tag{4}$$

where prompt p incorporates the desired narrative label n, difficulty value d, and target text t. The prompt follows this template:

> "Generate a $\langle d \rangle$ question-answer pair about narrative label $\langle n \rangle$ considering the following text: $\langle t \rangle$"

M is an encoder-decoder model that is fine-tuned using $D_{i-\text{augmented}} = (t, q, a, n, d)$ instances. The encoder receives prompt p and encodes it into a fixed-length representation known as a context vector. The decoder takes the context vector and generates the output text $(q_{\text{new}}, a_{\text{new}})$, using special tokens $\langle \text{QU} \rangle$ and $\langle \text{AN} \rangle$ that serve to differentiate between q_{new} and a_{new}. The idea is to guide the model in generating a question-answer pair of the intended difficulty d and narrative element n.

4 Experimental Setup

4.1 Preparing the FairytaleQA Dataset

We use FairyTaleQA (Xu et al., 2022), which comprises 10,580 question–answer pairs manually created by educational experts based on 278 narrative stories. Each story contains approximately 15 section texts, and each section (about 149 tokens) contains approximately 3 question–answer pairs. From the original dataset, we have prepared different data setups[2] for generating a QA pair:

- **Text → QA**: This setup only contains the text as input, so it serves as a baseline to compare with the subsequent setups, which consider control attributes.
- **Nar + Text → QA**: This setup considers *narrative* as a control attribute in the input.
- **Dif + Text → QA**: This setup considers *difficulty* as a control attribute in the input.
- **Nar + Dif + Text → QA**: This setup considers both the narrative and difficulty attributes.

4.2 Creating Simulated-Learner QA Systems

To create the simulated-learner QA systems, we trained five QA models. The choice of five was made empirically: it provided sufficient granularity for analysis while avoiding ties that could arise with fewer levels (e.g., four). The selected encoder models are `DeBERTaV3` (He et al., 2021), `RoBERTa` (Liu, 2019), `BERT` (Devlin et al., 2019) and `DistilBERT` (Sanh, 2019). We also use one decoder: `GPT-2` (Radford et al., 2019). They were fine-tuned on separate general-purpose question answering data (the SQuAD v1.1 dataset (Rajpurkar et al., 2016)). The models were deliberately chosen for their varying performance levels, thereby simulating different levels of learner skill. Table 1 shows the performance of each QA system on the SQuAD v1.1 evaluation set, using the n-gram similarity metric ROUGE_L-F1 (Lin, 2004) (QA answer vs. SQuAD ground-truth answer).

[2]The arrow separates the input (left) and output (right) information. On the left part, the + symbol illustrates whether the method incorporates control attributes.

Table 1: Simulated-Learner QA systems performance on SQuAD v1.1 evaluation set.

Sim.-Learner QA	ROUGE$_L$-F1 (0-1)
DeBERTaV3 (large)	0.87
RoBERTa (base)	0.82
BERT (base)	0.75
DistilBERT (base)	0.69
GPT-2	0.46

4.3 Answering FairytaleQA Questions with QA Systems

For each question in the train and validation sets of the FairyTaleQA dataset, all five simulated-learner QA systems generated their own answers. Each QA answer is then compared to the corresponding ground-truth answer to determine correctness. We considered an answer correct if it achieved either an exact match score of 1 or a ROUGE$_L$-F1 score of at least 0.5. The QA answers are organized into a binary response matrix — Figure 2 shows an example of such a matrix. Each row corresponds to a simulated-learner QA system and each column corresponds to a question ID. Each cell contains a 0 or 1, indicating incorrect or correct answers, respectively. This matrix serves as input data for the subsequent question difficulty estimation using IRT.

4.4 Estimating Question Difficulty with IRT

Based on the collected correct and incorrect answers for each question — organized into a binary response matrix — we estimated question difficulty using the Rasch Model (recall Section 2.2). Specifically, using the binary correctness data produced by the simulated-learner QA systems, the estimation is performed using the Expectation-Maximization (EM) algorithm (Embretson and Reise, 2000). This yielded difficulty values that were subsequently normalized to a 0-1 scale (0, 0.28, 0.50, 0.72, and 1), where higher values represent more difficult questions. The numerical values were converted into corresponding categorical labels – *easy*, *medium*, *moderate*, *hard*, and *extreme* – to be used in textual prompts. The distribution of the estimated difficulty values by narrative label in the data is presented in Table 2. Some attributes (e.g., *feeling* and *prediction*) have limited representation in the dataset.

Additionally, using the Maximum a Posteriori (MAP) algorithm (Embretson and Reise, 2000), we estimated the ability (θ) values for each QA system. These values are reported in Table 3, with higher values representing higher abilities. These values align, as expected, with the systems' original performance levels shown in Table 1.

Nar.	Easy	Med.	Mod.	Hard	Extr.
Action	773	362	375	435	749
Causal	316	200	245	316	1291
Char.	497	133	101	116	115
Feeling	55	79	62	89	539
Out.	126	114	138	165	268
Pred.	22	21	23	50	250
Setting	276	70	60	54	63
Action	76	40	65	60	92
Causal	35	27	31	50	151
Char.	50	17	14	9	17
Feeling	0	9	9	5	71
Out.	11	13	19	15	39
Pred.	1	3	6	7	38
Setting	29	4	5	4	3

Table 2: Difficulty values by Nar. (train and val set).

Sim.-Learner QA	Ability (θ)
DeBERTaV3 (large)	0.43
RoBERTa (base)	0
BERT (base)	-0.66
DistilBERT (base)	-1.25
GPT-2	-1.60

Table 3: Simulated-learner estimated ability values (θ) after answering questions from the FairytaleQA dataset.

We use *mirt*[3] tool for IRT, including all estimations.

4.5 Creating a Question-Answer Pair Generation Model

We use the Flan-T5 (Chung et al., 2024) encoder-decoder model for the controllable task. This model builds upon the original T5 (Raffel et al., 2020), which has been fine-tuned with task-specific instructions using prefixes, making it well-suited for our methodology. Additionally, Flan-T5

[3] https://cran.r-project.org/web/packages/mirt/index.html

demonstrates remarkable performance in text generation tasks, particularly in QG (Chen et al., 2024; Li and Zhang, 2024). We employ the `flan-t5-large` version, which is publicly available via Hugging Face[4]. Training is conducted for up to 10 epochs, with early stopping implemented using a patience of 2 epochs. During inference, we apply Top-k sampling with $k = 50$, $p = 0.9$ and $temp = 1.2$ to encourage diversity (values obtained experimentally). We initially explored beam search, a widely used technique in QG; however, we observed that it frequently produced repetitive questions when tasked with generating questions for the same narrative element across different difficulty levels.

4.6 Generating QA Pairs for Evaluation

We fine-tune the `Flan-T5` model on the training set of FairytaleQA. We obtain 4 models, as the model has been trained on each of the 4 data setups described in Section 4.1. For the 2 setups where difficulty labels are used, we apply the resulting models (inference) to the corresponding test set and generate 5 QA pairs for each text's section — one QA pair for each difficulty label. Since the FairytaleQA test set contains 394 section texts, we obtain a total of 1,970 generated QA pairs. Additionally, each text includes human-authored QA pairs associated with different narrative labels. This approach ensures that the generated QA pairs are balanced across distinct difficulty levels and narrative elements for further evaluation.

5 Evaluation

5.1 Evaluation Procedure

For NCQG, our evaluation protocol follows prior studies (Zhao et al., 2022; Leite and Lopes Cardoso, 2023, 2024) that focused on controlled generation using narrative labels. For DCQG, the evaluation protocol is based on recent works (Uto et al., 2023; Tomikawa et al., 2024; Tomikawa and Uto, 2024) that emphasize the use of simulated-learner QA systems across generated questions with distinct difficulty levels.

Narrative Control: To assess narrative control, we use a standard approach in QG: comparing generated questions directly with human-authored ground-truth questions. Hypothesis 1 (H1) is that *incorporating narrative attributes will result in generated questions that are more similar to the ground-truth*, as previously shown by Leite and Lopes Cardoso (2024). To quantify the similarity, we employ the *n*-gram similarity metric ROUGE$_L$-F1 (Lin, 2004), as originally adopted by the FairytaleQA authors. For a better perception of the idea, consider the human-authored ground-truth question: "What did Matte and Maie do on Saturdays?" (annotated with the *action* narrative element) and the generated question targeting the same narrative element: "What did Maie and Matte do to provide for themselves?". These questions yield a high ROUGE$_L$-F1 score because they are similar in terms of the narrative-related vocabulary they share, thus indicating successful narrative control.

Difficulty Control: For difficulty control, the evaluation focuses on analyzing the performance of simulated-learner QA systems when answering questions generated at varying difficulty levels. Hypothesis 2 (H2) posits that *simulated-learner QA systems will perform better on easier questions and worse on more difficult ones, relative to their ability levels*.

5.2 Results

Narrative Control: Table 4 presents the results from the narrative control perspective, measured using ROUGE$_L$-F1 *n*-gram similarity between the generated questions and the human-authored ground-truth questions. We observe an improvement in the similarity to ground-truth questions when narrative control attributes are incorporated. This trend is consistently observed across all seven narrative labels. Furthermore, these findings align with the results reported in prior studies on narrative control (Leite and Lopes Cardoso, 2023, 2024). Of novelty, when narrative and difficulty labels are fused, we observe a similar improvement trend, comparable to the incorporation of narrative attributes alone. These results support Hypothesis 1 (H1), indicating that our method effectively controls the narrative elements underlying the generated questions. Appendix A shows further support by reporting semantic similarity results.

Difficulty Control: Figure 3 presents the results for difficulty control only, showing the percentage of correct responses from the simulated-learner QA systems across all difficulty levels. The percentage of correct answers decreases as the difficulty level increases for all simulated learners[5]. Additionally,

[4] https://huggingface.co/google/flan-t5-large

[5] All percentages are relatively low (<60). This is because the QA models were not trained on the FairyTaleQA dataset but were instead trained on SQuAD. This intentional choice

Data Setup	Char.	Setting	Action	Feeling	Causal	Out.	Pred.
Text → QA	.227	.269	.287	.281	.271	.227	.251
Nar + Text → QA	.304	.537	.427	.527	.412	.458	.348
Nar + Dif + Text → QA	.305	.530	.412	.529	.405	.425	.365

Table 4: **Narrative Control**: Similarity (ROUGE$_L$-F1) between generated and ground-truth questions on the test set by narrative element. **Text → QA** is used as a baseline to assess whether narrative control helps the generated questions approximate the ground-truth questions.

learners with higher abilities achieve higher percentages of correct answers, while those with lower abilities achieve lower percentages. These findings are consistent with previous works (Uto et al., 2023; Tomikawa et al., 2024) and support Hypothesis 2 (H2), demonstrating that the method controls the difficulty levels of the generated questions.

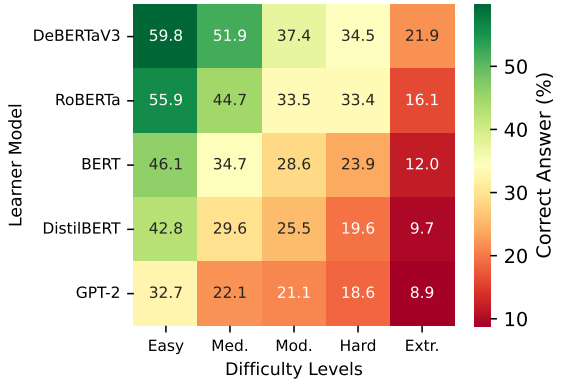

Figure 3: Percentage (%) of correct answers by difficulty level when only difficulty control labels are used (**Dif + Text → QA**).

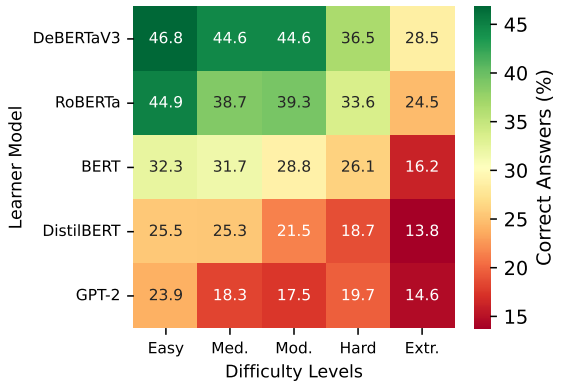

Figure 4: Percentage (%) of correct answers by difficulty level when both difficulty and narrative control labels are used (**Nar + Dif + Text → QA**).

ensures that the models' knowledge remains unbiased with respect to FairyTaleQA content.

Joint Narrative and Difficulty Control: Figure 4 presents the results for difficulty control when difficulty and narrative attributes are fused. In most cases, the percentage of correct answers decreases as the difficulty level increases across all simulated learners. These findings demonstrate that even when conditioning the generation process on both narrative content and difficulty, it remains possible to perform difficulty control. However, some inconsistencies are observed: for `DeBERTaV3`, there is no distinction between medium and moderate difficulty levels; for `RoBERTa`, the percentage of correct answers increases between medium and moderate levels; and for `GPT-2`, a similar trend occurs between moderate and hard levels. For an overall graphical comparison of difficulty control using only difficulty versus combining difficulty and narrative attributes, see Appendix B.

Figure 5 shows the overall accuracy for each narrative label, with trends suggesting difficulty control particularly between easy, hard, and extreme levels. However, control becomes inconsistent at intermediate levels. Among the attributes, *causal* and *outcome* demonstrate the most consistent control across difficulty levels, while *prediction* and *feeling* exhibit the least success. This inconsistency can be related to the limited representation of these attributes in the FairytaleQA dataset (recall Table 2), which prevents the model from learning to generate questions across different difficulty levels. Additionally, questions tied to these attributes are inherently more challenging, as reflected in the lower global performance of simulated-learner QA systems. For attributes such as *character*, *prediction*, *action*, and *setting*, the confusion is particularly evident between medium and moderate levels. To address this, we experimented with an alternative model trained on a lower granularity of difficulty levels, combining medium, moderate, and hard into a single medium level. In Figure 6, we show the result of this experiment, which demonstrates more consistent control across all levels.

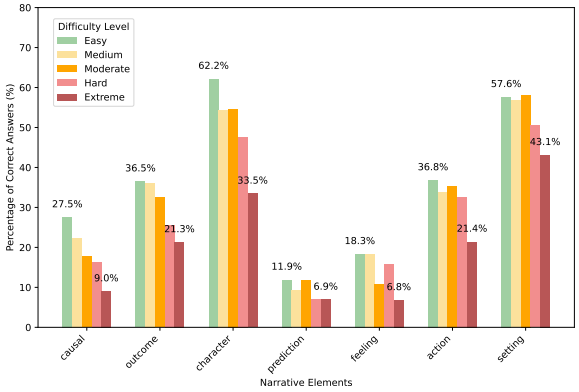

Figure 5: Percentage (%) of correct answers per narrative element and difficulty level (5 levels).

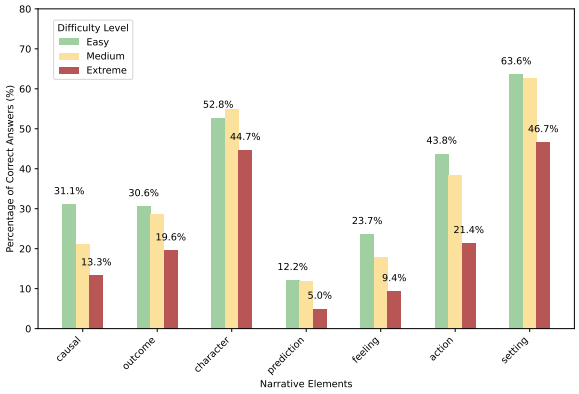

Figure 6: Percentage (%) of correct answers per narrative element and difficulty level (3 levels).

However, the *character* and *prediction* attributes continue to reveal some difficulty in distinguishing levels. These results support Hypothesis 2 (H2), confirming that the joint method enables difficulty control, although with less consistency than when controlling for difficulty alone. In Section 6, we outline potential explanations for these results.

Linguistic Features Influenced By Control: To better understand the linguistic features influenced by the controllability method, we analyze the linguistic properties of the generated QA pairs across different difficulty levels and narrative attributes. Prior work on difficulty-only controlled generation (Tomikawa et al., 2024) identifies two key factors that distinguish difficulty levels: (1) the average number of words in the generated answers, and (2) the distribution of initial interrogative terms in the generated questions. While we also explore these features (see Appendix C), we emphasize here a novel aspect that we also found experimentally to be relevant: (3) the degree of lexical novelty in the generated QA pairs relative to the source narrative text. To quantify this, we use the PINC (Paraphrase In N-gram Changes) metric (Chen and Dolan, 2011), which computes the percentage of n-grams present in the generated QA pairs but not in the source text. Higher PINC scores indicate greater lexical novelty and diversity. The results in Table 5 show that the diversity of the QA pairs increases with higher difficulty levels. This trend is observed both when difficulty labels are used independently and when combined with narrative labels. Therefore, we conclude that the linguistic diversity between the generated QA pairs and the source text is a feature influenced by difficulty control, regardless of whether difficulty labels are used alone or in conjunction with narrative labels.

Data Setup		Easy	Med.	Extr.
Dif + Text	Q	55.60	60.23	63.94
→ QA	A	9.88	23.17	48.69
Nar + Dif + Text	Q	57.34	60.72	65.57
→ QA	A	22.02	26.00	41.14

Table 5: PINC values (%) considering 3-gram novelty for generated questions (Q) and answers (A).

Error Analysis of Generated QA Pairs: We randomly selected 100 generated QA pairs and analyzed potential problems. Two main issues were identified, as shown in Figure 7: (1) *hallucinated content* (14%) and (2) *nonsensical QA pairs* (10%).

In the first example (1), the generated QA pair introduces made-up terms that do not exist in the source text. This is the most frequent problem encountered, likely due to the nature of the inference method (Top-k sampling). In example 2a, the problem is that the generated QA pair misleading. In Example 2b, the question seeks a motivation not in the text, and the vague answer just reuses the question's wording without adding meaning.

6 Discussion

We revisit our research question (RQ): *How effectively can we control the generation of question-answer pairs conditioned on both narrative and difficulty attributes using a modest scale model?*

Our findings indicate that the model effectively controls narrative elements, whether alone or combined with difficulty attributes. It also shows effective difficulty control when conditioned solely on difficulty attributes. The joint model, which combines narrative and difficulty attributes, generally achieves consistent control for at least three levels

Example 1 (**hallucinated content**):

Text: And then he cried for the third time: "Old Hopgiant!" And there appeared a fellow of truly horrible size, with a scythe as large as a ship's mast.
Gen. Q: Who appeared to the hunter at the swan?
Gen. A: A fellow of truly horrible size, with a scythe as large as a ship's mast.
Explanation: There is no mention of any hunter or swan in the source text.

Example 2a (**nonsensical QA pairs**):

Text: When the old man had finished his dance, the big demon said: "Thank you very much for your amusing dance. Now give us the pleasure of drinking a cup of wine with us," and with these words he handed him his largest wine-cup.
Gen. Q: What did the old man do after he finished dancing?
Gen. A: Danced with the big demon.
Explanation: The old man had finished dancing and did not dance with anyone again.

Example 2b (**nonsensical QA pairs**):

Text: But the rich neighbor wanted the meadow for himself alone, and told the poor one that he would drive him out of house and home if he did not come to an agreement with him that whichever one of them mowed the largest stretch of the meadowland in a single day, should receive the entire meadow.
Gen. Q: Why did the neighbor want to own the meadow?
Gen. A: He wanted to have the meadow.
Explanation: The question contains the answer.

Figure 7: Examples of problematic generated question-answer pairs (error analysis).

(easy, hard, and extreme). However, inconsistencies arise in the intermediate levels (medium and moderate). We also observed that certain attributes are more conducive to effective control, while others, like *prediction* and *feeling*, are less effective. Notably, reducing the granularity of difficulty levels improves the overall control. We now delve into two main factors that underlie our findings.

First, *generating QA pairs while simultaneously controlling both difficulty and narrative attributes is an inherently challenging task*. When the narrative element is fixed, the space of plausible questions becomes more constrained. This makes it harder to vary difficulty meaningfully, as the questions tend to focus on similar content. For instance, in Figure 1, the last two questions share the same narrative element but differ in difficulty. This overlap in content makes it harder to generate questions with clearly distinct difficulty levels.

Second, *some narrative attributes naturally lead to easier questions*. For instance, the *character* attribute often involves straightforward "Who" questions, making it harder to create questions with distinct difficulty levels. In contrast, questions following the *prediction* attribute are demanding, adding complexity to the learning process of generating well-differentiated questions.

Transferability to other domains: While our current work focuses on narrative comprehension, the principles of controllable QG are not domain-specific. For instance, it would be feasible to control generation based on other reading comprehension skills, as explored by Ghanem et al. (2022). Progress in this direction depends on the availability of datasets annotated with these dimensions, which are scarce.

Relevance to education: We believe our findings hold promise for educational applications, particularly in personalized QG. Recent work has explored adapting QG to student ability (Tomikawa et al., 2024). We argue that incorporating narrative control adds another valuable layer to personalization, enabling more targeted and contextually rich QG.

7 Conclusions

This work investigates a strategy for controlling both narrative and difficulty attributes in generated QA pairs. The results offer a preliminary yet promising demonstration of the potential of QG models and the proposed control strategy. Future efforts could leverage larger datasets with a more balanced distribution of questions across categories to improve the model's control capabilities. Additionally, examining the impact of different inference methods on generation would be valuable, especially to address the issue of repetitive outputs observed with beam search. Finally, future research could explore few-shot prompting techniques, providing minimal examples to assess the model's control ability without extensive training.

Limitations

While our approach provides promising insights into controllable QG, some limitations should be acknowledged.

First, *the limited representation of question categories across narrative attributes and difficulty levels hinders the model's ability to learn effectively*. FairytaleQA consists of approximately 10k instances. Associating questions with multiple narrative elements and difficulty levels significantly reduces the number of examples per category, limiting the model's ability to learn effectively. For instance, as shown previously in Table 2, *prediction* and *feeling* questions are poorly represented.

Second, *top-k sampling enables control over narrative elements and question difficulty but can lead to undesired hallucinations*. Initially, we experimented with beam search — a more commonly used technique for QG — but found it often generated repetitive questions when addressing the same narrative element across varying difficulty levels. Moreover, our findings indicate that the choice of inference method significantly impacts control. For instance, as shown in Section 5.2, the diversity of the generated QA pairs increases at higher difficulty levels. However, this diversity can also produce unintended side effects, such as the hallucinations noted with error analysis. While hallucinated QA pairs may affect evaluation by inflating perceived difficulty, we believe that reporting such cases was important to reveal potential failure modes of controllable QG systems. Although they may add some noise, these observations help contextualize the results and guide future improvements in model robustness.

Third, the *evaluation relies on simulated learner responses rather than real student data*. While this approach offers scalability and approximations of question difficulty, it may not fully reflect how actual students would respond. Nonetheless, it provides a valuable proxy for assessing the model's behavior, and we believe it still offers meaningful insight into the controllability of QG systems. Future work should explore incorporating real student data to further validate these findings.

Ethics Statement

This research involves the automatic generation of QA pairs from narrative texts, incorporating control attributes such as difficulty level and narrative elements. The dataset used, FairytaleQA, consists of human-authored QA pairs from publicly available fairy tales. No personally identifiable or sensitive information is included, ensuring compliance with ethical guidelines for data usage. The generated QA pairs were evaluated using both automatic metrics and manual inspection to identify potential errors, such as hallucinated content and nonsensical questions. We acknowledge that these models may introduce unintended errors or biases. While this paper does not focus on error mitigation, future work could explore extended human-in-the-loop validation to enhance the reliability of generated QA pairs, particularly in deployment scenarios.

Acknowledgments

The authors would like to thank Professor Masaki Uto and Yuto Tomikawa for their helpful clarifications and discussions related to prior work. This work was financially supported by UID/00027 – the Artificial Intelligence and Computer Science Laboratory (LIACC), funded by Fundação para a Ciência e a Tecnologia, I.P./ MCTES through national funds. Bernardo Leite is supported by a PhD studentship (with reference 2021.05432.BD), funded by FCT.

References

Julie Alonzo, Deni Basaraba, Gerald Tindal, and Ronald S Carriveau. 2009. They read, but how well do they understand? an empirical look at the nuances of measuring reading comprehension. *Assessment for Effective Intervention*, 35(1):34–44.

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Lizhen Qu, Shirong Shen, Guilin Qi, Lu Pan, and Yinlin Jiang. 2021. Simple or complex? complexity-controllable question generation with soft templates and deep mixture of experts model. *arXiv preprint arXiv:2110.06560*.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Jiaju Chen, Yuxuan Lu, Shao Zhang, Bingsheng Yao, Yuanzhe Dong, Ying Xu, Yunyao Li, Qianwen Wang, Dakuo Wang, and Yuling Sun. 2024. StorySparkQA: Expert-annotated QA pairs with real-world knowledge for children's story-based learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17351–17370, Miami, Florida, USA. Association for Computational Linguistics.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. 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 5968–5978, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie C. K. Cheung. 2023. How useful are educational questions generated by large language models? In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 536–542, Cham. Springer Nature Switzerland.

SE Embretson and SP Reise. 2000. Item response theory for psychologists. *Lawrence Earlbaum Associates, Mahwah, NJ*.

Yifan Gao, Lidong Bing, Wang Chen, Michael Lyu, and Irwin King. 2019. Difficulty controllable generation of reading comprehension questions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4968–4974. International Joint Conferences on Artificial Intelligence Organization.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer von der Ohe, and Alona Fyshe. 2022. Question generation for reading comprehension assessment by modeling how and what to ask. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2131–2146, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

David R. Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory Into Practice*, 41(4):212–218.

Vishwajeet Kumar, Yuncheng Hua, Ganesh Ramakrishnan, Guilin Qi, Lianli Gao, and Yuan-Fang Li. 2019. Difficulty-controllable multi-hop question generation from knowledge graphs. In *The Semantic Web – ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I*, page 382–398, Berlin, Heidelberg. Springer-Verlag.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Bernardo Leite and Henrique Lopes Cardoso. 2023. Towards enriched controllability for educational question generation. In *Artificial Intelligence in Education*, pages 786–791, Cham. Springer Nature Switzerland.

Bernardo Leite and Henrique Lopes Cardoso. 2024. On few-shot prompting for controllable question-answer generation in narrative comprehension. In *Proceedings of the 16th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 63–74. INSTICC, SciTePress.

Kunze Li and Yu Zhang. 2024. Planning first, question second: An LLM-guided method for controllable question generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4715–4729, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.

Julie S Lynch, Paul Van Den Broek, Kathleen E Kremer, Panayiota Kendeou, Mary Jane White, and Elizabeth P Lorch. 2008. The development of narrative comprehension and its relation to other early reading skills. *Reading Psychology*, 29(4):327–365.

Alison H Paris and Scott G Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1):36–76.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Vasile Rus, Zhiqiang Cai, and Art Graesser. 2008. Question generation: Example of a multi-year evaluation campaign. *Proc WS on the Question Generation Shared Task and Evaluation Challenge*.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Yuto Tomikawa, Ayaka Suzuki, and Masaki Uto. 2024. Adaptive question–answer generation with difficulty control using item response theory and pretrained transformer models. *IEEE Transactions on Learning Technologies*, 17:2240–2252.

Yuto Tomikawa and Masaki Uto. 2024. Difficulty-controllable multiple-choice question generation for reading comprehension using item response theory. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 312–320, Cham. Springer Nature Switzerland.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. Generative language models for paragraph-level question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 670–688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 119–129, Toronto, Canada. Association for Computational Linguistics.

Xu Wang, Simin Fan, Jessica Houghton, and Lu Wang. 2022a. Towards process-oriented, modular, and versatile question generation that meets educational needs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 291–302, Seattle, United States. Association for Computational Linguistics.

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G. Baraniuk. 2022b. Towards human-like educational question generation with large language models. In *Artificial Intelligence in Education*, pages 153–166, Cham. Springer International Publishing.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Trans. Inf. Syst.*, 40(1).

Sanqiang Zhao, Seokhwan Kim, Yang Liu, Robinson Piramuthu, and Dilek Hakkani-Tür. 2023. Storyqa: Story grounded question answering dataset.

Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5073–5085, Dublin, Ireland.

A Narrative Control: Semantic Similarity

Table 6 presents the results from the narrative control perspective, measured using BLEURT (Sellam et al., 2020). The goal is to show an improvement in semantic similarity to ground-truth questions when narrative control attributes are incorporated. As observed with ROUGE$_L$-F1 similarity (recall Section 5.2), this trend is observed across all seven narrative labels. When narrative and difficulty labels are fused, we observe a similar improvement trend, comparable to the incorporation of narrative attributes alone. These results further support Hypothesis 1 (H1) — *incorporating narrative attributes will result in generated questions that are more similar to the ground-truth* — indicating that our method controls the narrative elements underlying the generated questions.

B Difficulty-Only vs. Difficulty+Narrative Control

To compare difficulty control when operating solely on difficulty versus combining difficulty and narrative attributes, Figure 8 provides an overview of the performance at each level for both setups. Both setups show the expected trend: the percentage of correct answers decreases as difficulty increases. However, a linear approximation of the observed data points reveals that the decrease is less pronounced when both attributes are combined, though it remains consistent overall.

Data Setup	Char.	Setting	Action	Feeling	Causal	Out.	Pred.
Text → QA	.332	.332	.353	.370	.360	.346	.358
Nar + Text → QA	.379	.504	.422	.491	.418	.444	.409
Nar + Dif + Text → QA	.378	.482	.413	.499	.417	.422	.401

Table 6: **Narrative Control**: Semantic similarity (BLEURT) between generated and ground-truth questions on the test set by narrative element. **Text → QA** is used as a baseline to assess whether narrative control helps the generated questions approximate the ground-truth questions.

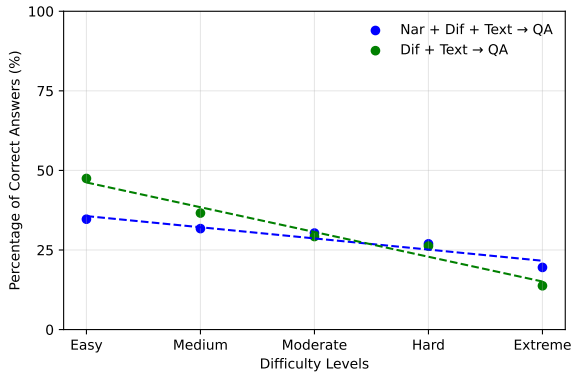

Figure 8: Percentage of Correct Answers by Dif. Level.

C Additional Linguistic Features Influenced By Control

Table 7 presents the average number of words in the generated question-answer pairs. For generated answers, when only difficulty labels are incorporated, no significant trend is observed. For generated questions, an upward trend is noted, though it is not significant. When narrative and difficulty labels are combined, no trend is observed. Based on these findings, we conclude that the average length of generated question-answer pairs is not influenced by difficulty or narrative control labels in our experiments.

Data Setup		Easy	Med.	Extr.
Dif + Text	Q	10.80	11.83	12.49
→ QA	A	7.19	8.95	8.88
Nar + Dif + Text	Q	11.81	11.62	11.70
→ QA	A	7.42	7.96	7.61

Table 7: Average number of words for generated questions (Q) and answers (A).

Figure 9 illustrates the proportion of initial interrogative terms in the generated questions. When only difficulty labels are used (top chart), higher difficulty levels show an increase in terms like "why" and "how" and a decrease in terms like "what" "who" and "where". This aligns with expectations, as "why" and "how" are often linked to questions requiring higher cognitive effort, as described in Bloom's taxonomy (Krathwohl, 2002). When both narrative and difficulty labels are fused (lower chart), the proportion of all interrogative terms is more consistent across difficulty levels. This outcome is expected since this setup aims to control difficulty levels while also demanding for certain narrative elements. In this case, narrative labels are the primarily influence for the choice of interrogative terms (e.g., "who" for character-related questions), rather than difficulty labels.

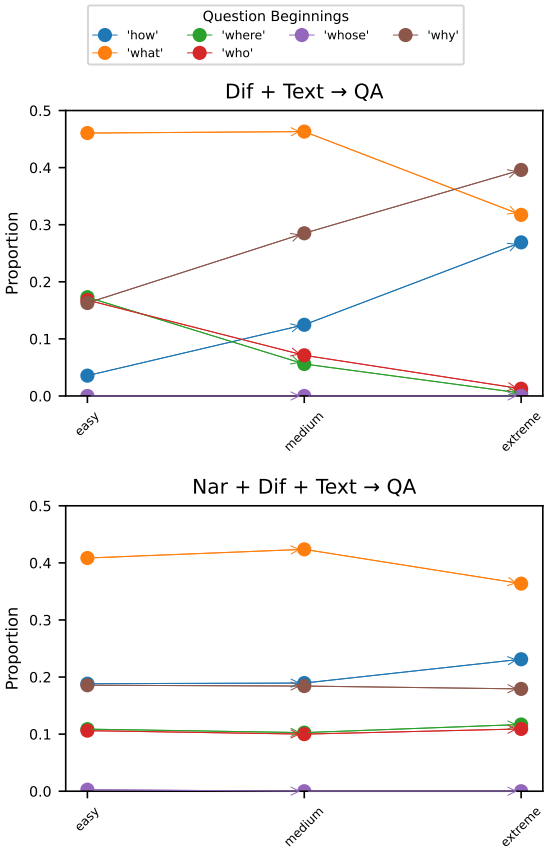

Figure 9: Proportion of initial interrogative terms in the generated questions (arrowed lines indicate increase/decrease trends).