通过联合叙述和难度控制推进问题生成

Preprint. Accepted to the BEA 2025 Workshop (ACL)

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

问题生成（QG）是从源输入自动生成问题的任务，近年来取得了显著进展。难度可控问题生成（DCQG）在考虑学习者能力的同时，能够控制生成问题的难度级别。此外，叙事可控问题生成（NCQG）允许对问题中嵌入的叙事方面进行控制。然而，问题生成研究缺乏对结合这两种控制类型的关注，这对于生成适合教育目的的问题非常重要。为了解决这一空白，我们提出了一种联合叙事和难度控制的策略，使得在阅读理解问题的生成中能够同时控制这两个属性。我们的评估提供了这一方法可行的初步证据，尽管它在所有实例中并不有效。我们的研究结果突出了该策略表现良好的条件，并讨论了其应用所涉及的权衡。

1 介绍

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 : "简单"
Generated QA Pair : Who challenged the turtle to a race?" The hare

---

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

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

Figure 1: 带有不同难度等级和叙述属性的受控问答生成的示例。

问题生成（QG）专注于针对数据源（包括非结构化文本或知识库 (**?**) ）的自动生成连贯且有意义的问题。可控问题生成在教育 (**?**) 中起着至关重要的作用，因为它有助于生成个性化的问题，以满足学生独特的需求和学习目标。最近关于 QG 的工作利用了微调 (**??**) 和少样本提示 (**??**) 等技术，根据源文本和可选的目标答案生成问题。在可控 QG 中，通过在输入或提示中加入可控标签以指导生成过程来增强这个过程。具体来说，关于叙事控制问题生成（NCQG）的研究集中于通过潜在的叙事元素（例如，因果关系）(**???**) 来指导生成问题的内容。反过来，难度可控问题生成（DCQG）强调控制回答问题的预期难度 (**????**) 。一些研究考虑了问题难度与学习者能力之间的关系 (**??**) 。然而，在可控 QG 研究中缺乏这两种控制方式的结合，这对于在这一领域生成模型日益增加的使用中促进人类控制尤其重要 (**?**) 。因此，本研究提出了一种策略，探索将叙事和难度控制结合起来的可行性，以生成面向儿童的叙事故事的阅读理解问答（QA）对。图 1 展示了该策略的一个示例。我们正式研究以下研究问题（RQ）：在多大程度上我们能够有效地控制基于叙事和难度属性的问答对的生成，使用一个适度大小的 [1] 模型？在我们的实验中，我们使用了一个著名的数据集——FairyTaleQA (**?**) ——其中每个问题都已经标注了七个叙事标签之一。我们的方法包括两个主要步骤：（1）使用模拟学习者 QA 系统回答 FairyTaleQA 的问题，从而通过项目反应理论估计难度标签，以及（2）应用联合的叙事和难度控制模型，利用人工标注的叙事标签和为每个问题估计的难度标签。评估所提出的方法以确定是否成功地将 NCQG 和 DCQG 应用于生成的问题。对于 NCQG，我们比较人工编写的问题和生成的问题之间的相似性。对于 DCQG，我们评估模拟学习者 QA 系统在生成的具有不同难度级别问题上的性能。尽管结果显示策略的有效性，NCQG 表现出一致的成功，而 DCQG 表现出中等成功，性能在特定的叙事属性和难度水平之间有所不同。我们的目标是强调这一策略在何种条件下表现出高效或低效，为从事类似研究的研究人员提供见解。总之，我们的贡献是：

[1] 超过 10 亿个参数。

- 我们提出了一种联合策略，用于在叙述和难度属性的条件下控制问答对的生成。
- 我们报告了受控制影响的语言特征，并对生成的问答对进行了错误分析，提供了对该方法性能和局限性的洞察。

2 背景和相关工作

2.1 可控问题生成 (CQG)

如 **?** 所述，先前关于 CQG 的研究主要探索了两个方面：内容（或类型）和难度。内容控制与融入生成问题中的语言元素有关。例如，**?** 提出控制特定的阅读理解技能，如修辞语言和词汇。此外，**?** 专注于叙述元素的控制，尽管 **?** 通过控制显性属性扩展了这种方法。**?** 建议控制布鲁姆的问题分类法 (**?**) 。

难度控制与回答生成问题的挑战有关，这一概念通常是主观的（即，难度可能会因回答者不同而有所变化）。在这方面，**?** 基于 QA 系统是否能够正确回答问题，为问题分配了难度标签（简单或困难），并使用这些标签作为输入来控制生成过程。**?** 提出了基于命名实体流行度来估计难度的方法，而 **?** 解决了 QG 中高多样性的挑战。此外，**?** 通过考虑得到答案所需的推理步骤数量来控制问题的难度。

以前方法的一个限制是（1）缺乏对问题难度与学习者能力之间关系的重视。为了解决这个问题，**?** 提出了使用项目反应理论（IRT）(**?**)，这是测试理论中的一个数学框架，用于量化问题难度并将其直接与学习者能力关联起来。另一个限制是（2）缺乏多属性的整合。虽然 **?** 结合了叙述和难度属性，他们在回答显著性及回答问题所需的句子数量来定义难度。本研究的新颖之处在于通过叙述元素来整合内容控制，与通过模拟学习者的能力得到的难度控制相结合，从而在之前研究的基础上进行构建。

2.2 项目反应理论 (IRT)

IRT (**?**) 是一个用于研究考生（能力或熟练程度）与其在测试项目上的表现之间相互作用的统计框架。IRT 的一个关键方面是建模问题难度与学习者能力之间的关系，从而提供关于一个问题如何区分具有不同技能水平的个体的见解。这种关系允许估计具有特定能力水平的学习者能正确回答给定问题的可能性，使其对自适应测试和理解问题的复杂性特别有用。在 IRT 中常用的模型是 Rasch 模型，它假设正确回答的概率取决于学习者的能力（θ）与项目难度（b）之间的关系：

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

，其中 θ_i 是个体 i 的学习能力，b_j 是项目 j 的难度，$P(X_{ij} = 1 \mid \theta_i, b_j)$ 是个体 i 正确回答项目 j 的概率。在我们的研究中，我们使用 IRT 来估计问题难度（b）和学习者能力（θ）参数。

2.3 FairyTaleQA：目的与价值

我们使用 FairytaleQA 数据集 (**?**)，因为其故事及相应的问答对符合解决叙事理解的目标。根据 **?**，叙事理解代表了一种与整体阅读能力密切相关的高级认知技能 (**?**) 。FairytaleQA 的一个关键特点是每个问题的专家注释，这些注释建立在基于证据的框架上 (**??**) 。被标注用于控制的叙事元素包括：

- 设定：聚焦于事件的时间和地点，通常以“在哪里……?”或“什么时候……”开始；
- 动作：与角色的行为相关；
- 感觉：探索情绪状态或反应（例如，“X 感觉如何?”）；
- 因果关系：解决因果问题（例如，“为什么……?”或“是什么导致/引发了 X?”）；
- 结果解析：关注事件的结果（例如，“X 之后发生了什么/会发生什么?”）；
- 预测：基于文本证据对未来或未知事件的提问。

虽然还有其他受欢迎的教育问答数据集（遵循开放式 wh-问题格式），如 NarrativeQA (**?**) 和 StoryQA (**?**)，但它们没有标注具体的阅读理解技能。这进一步促使我们决定在本研究中使用 FairytaleQA。

3 方法

本节概述了本研究的方法，包括添加基于 IRT 的困难标签给 FairytaleQA，并开发一个具有叙事和难度控制功能的问题-答案对生成模型。图 2 提供了本节讨论步骤的概览。

3.1 通过基于 IRT 的问题难度标签扩充 FairytaleQA

令 D 为我们的数据集，该数据集由以下四元组表示的实例组成：

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

，其中 t 是文本，q 是问题，a 是关于文本的答案，n 是与问题答案对 (q, a) 相关的叙述元素。目标是创建第五个元素 d，从而生成一个新的增强实例：

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

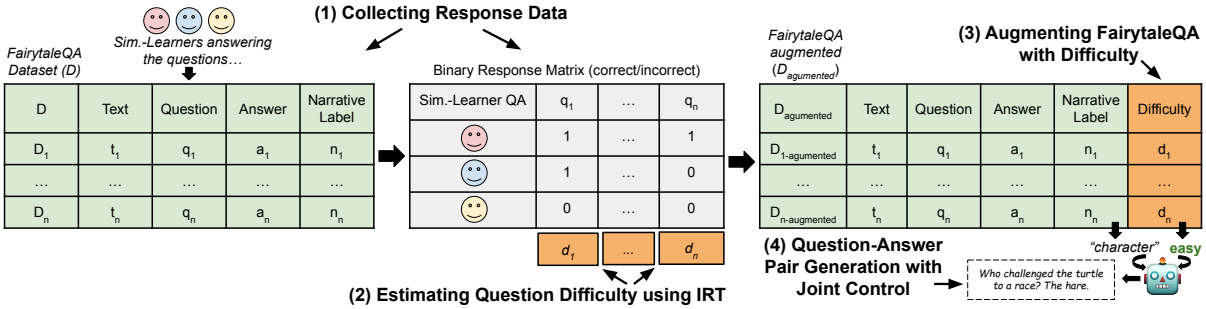

Figure 2: 联合叙述与难度控制的整体方法。

，其中 d 是与问题答案对 (q, a) 相关联的预计难度值。为了创建这些增强实例，我们采用了由 **?** 和 **?** 提出的方法：

1. 收集每个问答对的响应数据：我们从多个回答者那里收集了对问题的答案。鉴于缺乏真正的学生，我们使用了模拟学习者问答系统，这些模型能够自动提取所提出问题的答案。如第 4.2 节所述，问答模型是故意选择的，以代表不同的性能水平，从而模拟不同的能力水平。

2. 使用 IRT 估计问题难度：我们使用从模拟学习者问答系统收集的答案，通过 IRT 估计每个问题的难度，具体来说采用了在第 2.2 节中描述的 Rasch 模型。

3. 使用难度估计值扩充 FairytaleQA 数据集：基于估计的难度值，我们用 d 扩充数据集中的每个实例，最终结果为 $D_{i-\text{augmented}} = (t, q, a, n, d)$ 。

。

3.2 带有联合叙述和难度控制的问题-答案对生成

可控过程可以表示如下：给定指令提示 p ，目标是使用模型 M 生成一个问题-答案对 $(q_{\text{new}}, a_{\text{new}})$ 。这可以被公式化为：

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

其中提示 p 包含所需的叙述标签 n ，难度值 d ，和目标文本 t 。提示遵循以下模板：

> “生成一个关于叙述标签 $\langle n \rangle$ 的 $\langle d \rangle$ 问答对，考虑以下文本：$\langle t \rangle$ ”

M 是一个经过微调的编码-解码模型，使用 $D_{i-\text{augmented}} = (t, q, a, n, d)$ 实例。编码器接收提示 p 并将其编码为一个固定长度的表示，称为上下文向量。解码器采用上下文向量生成输出文本 $(q_{\text{new}}, a_{\text{new}})$ ，使用特殊标记 〈 QU 〉 和 〈 AN 〉 来区分 q_{new} 和 a_{new} 。其目的是引导模型生成预期难度的 d 和叙述元素 n 的问题-答案对。

4 实验设置

4.1 准备 FairytaleQA 数据集

我们使用 FairyTaleQA (**?**) ，该数据集由教育专家基于 278 个叙述故事手动创建的 10,580 对问答对组成。每个故事包含大约 15 个部分文本，每个部分（约 149 个标记）包含大约 3 个问答对。基于原始数据集，我们准备了不同的数据设置 [2] 用于生成一个问答对：

- 文本 → QA：此设置仅包含文本作为输入，因此它作为基准用于与后续考虑控制属性的设置进行比较。
- Nar + 文本 → QA：这种设置将叙述视为输入中的控件属性。
- Dif + 文本 → QA：此设置认为难度是输入中的一个控制属性。
- Nar + Dif + Text → QA：这个设置考虑了叙述和难度属性。

4.2 创建模拟学习者问答系统

为了创建模拟学习者的问答系统，我们训练了五个问答模型。选择五个模型是基于经验的：它提供了足够的分析粒度，同时避免了层次较少（例如，四个）可能会出现的并列情况。选定的编码器模型是 `DeBERTaV3` (**?**) 、`RoBERTa` (**?**) 、`BERT` (**?**) 和 `DistilBERT` (**?**) 。我们还使用了一个解码器：`GPT-2` (**?**) 。这些模型在独立的一般问答数据集（SQuAD v1.1 数据集 (**?**) ）上进行了微调。模型的选择是有意为之，以其不同的性能水平，从而模拟不同水平的学习者技能。表 1 展示了每个问答系统在 SQuAD v1.1 评估集上的性能，使用 n-gram 相似度指标 ROUGE

[2] 箭头将输入（左）和输出（右）信息分开。在左侧部分，+ 符号说明该方法是否包含控制属性。

$_L$ -F1 (**?**)（问答答案与 SQuAD 标准答案进行比较）。

Table 1: 模拟学习者 QA 系统在 SQuAD v1.1 评估集上的表现。

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 使用问答系统回答 FairytaleQA 问题

对于 FairyTaleQA 数据集的训练集和验证集中的每个问题，所有五个模拟学习者 QA 系统都生成了各自的答案。然后将每个 QA 答案与对应的真实答案进行比较，以确定正确性。如果答案获得了完全匹配分数 1 或 ROUGE $_L$ -F1 得分至少为 0.5，则认为答案是正确的。QA 答案被组织成一个二进制响应矩阵————图 2 展示了这样一个矩阵的示例。每一行对应一个模拟学习者 QA 系统，每一列对应一个问题 ID。每个单元格包含 0 或 1，分别表示错误或正确答案。此矩阵作为后续使用 IRT 进行问题难度估计的输入数据。

4.4 用 IRT 估计问题难度

基于每个问题收集的正确和错误答案——组织成一个二进制响应矩阵——我们使用 Rasch 模型估计了问题难度（回顾第 2.2 节）。具体而言，使用由模拟学习者问答系统生成的二进制正确性数据，通过期望最大化（EM）算法 (**?**) 进行估计。这产生了随后归一化到 0-1 比例（0, 0.28, 0.50, 0.72, 和 1）的难度值，其中更高的值表示更难的问题。这些数值被转换为对应的分类标签——简单、中等、一般、困难和极难——用于文本提示。数据中叙述标签估计难度值的分布展示在表 2 中。某些属性（例如，感觉和预测）在数据集中仅有有限的表示。

此外，使用最大后验概率（MAP）算法 (**?**)，我们估计了每个 QA 系统的能力（θ）值。这些值在表 3 中报告，较高的值代表较高的能力。正如预期的那样，这些值与表 1 中显示的系统原始性能水平相一致。

我们使用 mirt [3] 工具进行 IRT，包括所有估算。

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

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: 由 Nar 确定的难度值。（训练和验证集）。

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: 在回答 FairytaleQA 数据集的问题后，模拟学习者估计的能力值 (θ)。

4.5 创建一个问答对生成模型

我们使用 Flan-T5 (**?**) 编码-解码器模型来处理可控任务。该模型基于原始的 T5 (**?**) 进行构建，经过使用前缀的任务特定指令微调，因此非常适合我们的方法。此外，Flan-T5 在文本生成任务中表现出色，特别是在 QG (**??**) 方面。我们采用了 flan-t5-large 版本，可以通过 Hugging Face [4] 公开获取。训练进行至多 10 个 epoch，并采用具有 2 个 epoch 耐心度的提前停止策略。在推理过程中，我们应用 Top-k 采样结合 $k = 50$ 、$p = 0.9$ 和 $temp = 1.2$ 以鼓励多样性（这些值通过实验获得）。我们最初探索了波束搜索，这是一种在 QG 中广泛使用的技术；然而，我们发现当它负责为同一叙述元素生成不同难度等级的问题时，它频繁地产生重复的问题。

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

4.6 生成用于评估的问答对

我们在 FairytaleQA 的训练集上微调了 Flan-T5 模型。我们获得了 4 个模型，因为每一种数据设置都在 4.1 节中描述的 4 种数据设置上进行了训练。对于使用难度标签的 2 种设置，我们将得到的模型（推理）应用于相应的测试集，并为每个文本的部分生成 5 个问答对——每个难度标签一个问答对。由于 FairytaleQA 测试集包含 394 个部分文本，我们总共获得 1970 个生成的问答对。此外，每个文本还包括与不同叙事标签相关的人类创作的问答对。这种方法确保生成的问答对在进一步评估中在不同难度水平和叙事元素之间保持平衡。

5 评估

5.1 评估程序

对于 NCQG，我们的评估协议遵循以前的研究 (**???**)，这些研究集中于使用叙事标签进行的可控生成。对于 DCQG，评估协议基于最近的工作 (**???**)，这些工作强调在生成的具有不同难度级别的问题中使用模拟学习者问答系统。叙事控制：为了评估叙事控制，我们使用 QG 中的标准方法：将生成的问题与人类创作的真实问题直接比较。假设 1（H1）是，整合叙事属性将生成更类似于真实问题的问题，正如之前由 **?** 所示。为了量化相似性，我们采用了由 FairytaleQA 作者最初采用的 n-gram 相似性度量标准 ROUGE $_L$ -F1 (**?**)。为了更好地理解这个概念，考虑人类创作的真实问题："Matte 和 Maie 在星期六做了什么?"（用场景叙事元素标注）以及针对相同叙事元素的生成问题："Maie 和 Matte 做了什么来养活自己?" 这些问题因为在共享的叙事相关词汇方面相似，从而产生了较高的 ROUGE $_L$ -F1 得分，这表明了成功的叙事控制。

难度控制：对于难度控制，评估的重点是分析模拟学习者问答系统在回答不同难度级别生成的问题时的表现。假设 2 (H2) 提出，相对于它们的能力水平，模拟学习者问答系统在回答简单问题时表现会更好，而在回答困难问题时表现会更差。

5.2 结果

叙述控制：表格 4 显示了叙述控制视角的结果，通过生成的问题与人工编写的真实问题之间的 ROUGE $_L$ -F1 n -gram 相似性进行衡量。我们观察到，当融入叙述控制属性时，与真实问题的相似性有所提升。这个趋势在所有七个叙述标签中都一致地被观察到。此外，这些发现与之前关于叙述控制 (**??**) 的研究结果一致。有新意的是，当叙述和难度标签融合时，我们观察到相似的改进趋势，这与仅仅加入叙述属性的情况相当。这些结果支持假设 1 (H1)，表明我们的方法能有效控制生成问题的叙述元素。附录 A 通过报告语义接近度结果进一步支持这一点。

难度控制：图 3 仅展示了难度控制的结果，显示了模拟学习者问答系统在所有难度级别上的正确应答百分比。对于所有模拟学习者，正确答案的百分比随着难度水平的增加而下降。此外，能力较高的学习者获得更高的正确答案百分比，而能力较低的学习者获得较低的百分比。这些发现与之前的研究 (**??**) 一致，并支持假设 2（H2），证明该方法能够控制生成问题的难度级别。

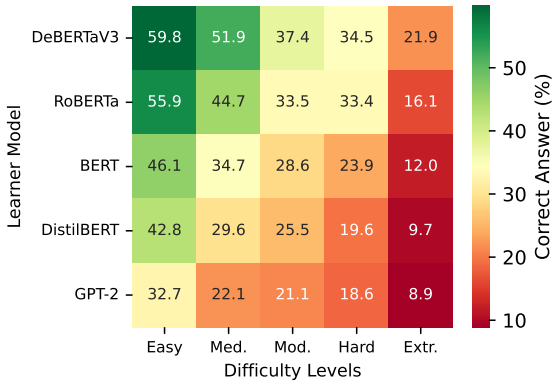

Figure 3: 仅使用难度控制标签时（Dif + Text → QA），按难度级别划分的正确答案百分比 (%)。

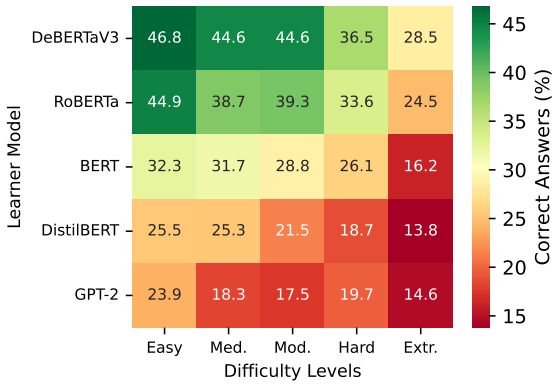

Figure 4: 当使用难度和叙述控制标签时，不同难度级别的正确答案的百分比 (Nar + Dif + Text → QA)。

联合叙述和难度控制：图 4 展示了在难度和叙述属性融合时进行难度控制的结果。在大多数情况下，随着所有模拟学习者难度等级的增加，正确答案的百分比都会下降。这些发现表明，即使在生成过程中同时考虑叙述内容和难度，也仍然可以进行难度控制。然而，观察

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: 叙述控制：测试集上生成的问题与真实问题之间的相似性（ROUGE $_L$ -F1），按叙述元素分类。将文本 → QA 作为基线，用于评估叙述控制是否有助于生成的问题接近真实问题。

到一些不一致：对于 DeBERTaV3，中等和适中难度等级之间没有区别；对于 RoBERTa，在中等和适中等级之间正确答案的百分比增加；对于 GPT-2，在适中和困难等级之间也出现了类似的趋势。关于仅使用难度与结合难度和叙述属性进行难度控制的整体图形比较，请参见附录 B。

图 **??** 显示了每个叙述标签的整体准确性，趋势表明在简单、困难和极端级别之间的难度控制尤其困难。然而，在中间级别的控制变得不一致。在属性中，因果和结果在不同难度级别中表现出最一致的控制，而预测和感受的成功率最低。这种不一致可能与 FairytaleQA 数据集中这些属性的有限表示有关（回顾表格 2），这使得模型难以学习在不同难度级别生成问题。此外，与这些属性相关的问题本质上更具挑战性，这反映在模拟学习者 QA 系统的较低整体性能中。对于诸如角色、预测、动作和背景等属性，混淆在中等和适度级别之间尤为明显。为了解决这个问题，我们尝试了一个在较低难度级别颗粒度上训练的替代模型，将中等、适度和困难结合为一个中等级别。在图 **??** 中，我们展示了该实验的结果，证明了在所有级别上更加一致的控制。然而，角色和预测属性在区分级别方面仍然显示出一些困难。这些结果支持假设 2 (H2)，确认联合方法能够实现难度控制，但比仅控制难度时的一致性稍差。在第 6 节中，我们概述了对这些结果的潜在解释。

受控影响的语言特征：为了更好地理解可控方法对语言特征的影响，我们分析了在不同难度级别和叙述属性下生成的问答对的语言属性。先前关于仅难度控制生成的研究 (**?**) 识别出区分难度级别的两个关键因素：（1）生成答案的平均字数，以及（2）生成问题中初始疑问词的分布。虽然我们也探讨了这些特征（见附录 C），但我们在此强调一个我们通过实验发现的相关新方面：（3）生成的问答对相对于源叙述文本的词汇新颖程度。为了量化这一点，我们使用了 PINC（短语在 N-gram 变化中）的指标 (**?**)，该指标计算生成问答对中存在但源文本中不存在的 n-gram 的百分比。较高的 PINC 分数表示更高的词汇新颖性和多样性。表格 5 的结果显示问答对的多样性随着难度级别的提高而增加。这一趋势在仅使用难度标签以及结合叙事标签使用时均有所观察。因此，我们得出结论，无论难度标签是单独使用还是与叙事标签结合使用，生成的问答对与源文本之间的语言多样性是受难度控制影响的特征。

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: 考虑到生成的问题 (Q) 和答案 (A)，3-gram 新颖性的 PINC 值 (%)。

生成的问答对的错误分析：我们随机选择了 100 个生成的问答对并分析了潜在的问题。发现了两个主要问题，如图 5 所示：(1) 虚假的内容（14 %）和 (2) 没有意义的问答对（10 %）。在第一个例子 (1) 中，生成的问答对引入了在原始文本中不存在的编造术语。这是最常遇到的问题，可能是由于推理方法（Top-k 抽样）的性质。在例子 2a 中，问题在于生成的问答对具有误导性。在例子 2b 中，问题寻求文本中不存在的动机，而模糊的答案只是重复问题的措词，没有增加意义。

6 讨论

我们重新审视我们的研究问题（RQ）：我们能多大程度上通过一个适度规模的模型控制以叙述和难度属性为条件的问题-答案对的生成？我们的研究结果表明，该模型能够有效控制叙述要素，无论是单独还是与难度属性结合。当仅以难度属性为条件时，也展示了有效的难度控制。结合叙述和难度属性的联合模型通常能够在至少三个级别（简单、困难和极端）上实现一致控制。然而，在中间级别（中等和适度）会出现不一致的情况。我们还观察到，某些属性更有利于有效控制，而其他属性，如预测和感觉，则不那么有效。值得注意的是，降低难度级别的粒度改善了整体控制。现在我们深入探讨导致我们研究结果的两个主要因素。

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 猎人 at the 天鹅?
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 : 与大魔共舞.
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 拥有草地?
Gen. A : He wanted to 拥有草地.
Explanation : The question contains the answer.

Figure 5: 生成的问题-答案对中的问题示例（错误分析）。

首先，同时控制难度和叙事属性来生成问答对本质上是一个具有挑战性的任务。当叙事元素固定时，合理问题的空间变得更加受限。这使得难以有意义地变化难度，因为问题往往集中在相似的内容上。例如，在图 1 中，最后两个问题共享相同的叙事元素但难度不同。这种内容上的重叠使得生成具有明显不同难度的问题更加困难。

其次，一些叙述属性自然会导致更简单的问题。例如，角色属性通常涉及简单的“谁”的问题，使得创建具有不同难度级别的问题更困难。相比之下，遵循预测属性的问题要求高，为生成具有良好区分度的问题的学习过程增加了复杂性。向其他领域的可转移性：虽然我们目前的工作集中在叙述理解上，但可控的问答生成的原则并不限于特定领域。例如，可以根据其他阅读理解技能来控制生成，如 **?** 所探索的那样。在这一方向上的进展取决于用这些维度注释的数据集的可用性，而这些数据集目前相对稀缺。与教育的相关性：我们相信我们的发现对教育应用尤其是在个性化问答生成方面有希望。最近的工作探索了将问答生成适应学生能力 (**?**) 。我们认为，结合叙述控制为个性化添加了另一种有价值的层次，使更具针对性和情境丰富的问答生成成为可能。

7 结论

这项工作探讨了一种策略，用于控制生成的问答对中的叙述和难度属性。结果提供了一个初步但有希望的展示，展示了 QG 模型及其所提出的控制策略的潜力。未来的工作可以利用更大的数据集，而这些数据集在各个类别的问题上具有更均衡的分布，以提高模型的控制能力。另外，研究不同推理方法对生成的影响也很有价值，尤其是解决使用束搜索时观察到的重复输出问题。最后，未来的研究可以探索少样本提示技术，提供极少的例子来评估模型在不进行大量训练的情况下的控制能力。

8

局限性

尽管我们的方法为可控的 QG 提供了有希望的见解，但一些局限性仍需被承认。

首先，叙事属性和难度级别的有限问题类别表示限制了模型的有效学习能力。FairytaleQA 大约包含 1 万个实例。将问题与多种叙事元素和难度级别关联会显著减少每个类别的示例数量，从而限制模型的有效学习能力。例如，如之前表 2 所示，预测和感觉问题的表示较差。

其次，top-k 抽样可以控制叙述元素和问题难度，但可能导致不良的幻觉现象。最初，我们尝试了 beam search——一种更加常用的 QG 技术——但发现它在对不同难度级别的同一个叙述元素进行处理时，常常产生重复性的问题。此外，我们的结果表明，选择推理方法对控制效果有显著影响。例如，如第 5.2 节所示，在较高难度级别时生成的问答对的多样性增加。然而，这种多样性也可能产生意外的副作用，例如在错误分析中注意到的幻想。尽管幻觉的问答对可能通过提升感知难度影响评估，但我们认为报告此类情况对于揭示可控 QG 系统的潜在失效模式是重要的。尽管它们可能增加一些噪声，这些观察帮助对结果进行背景化并指导未来模型稳健性的改进。第三，评估依赖于模拟的学习者响应而非真实的学生数据。虽然这种方法提供了可扩展性和问题难度的近似，但可能无法完全反映实际学生的响应方式。然而，它提供了一个有价值的代理来评估模型的行为，我们认为它仍然为 QG 系统的可控性提供了有意义的洞察。未来的工作应探索结合真实学生数据以进一步验证这些发现。

9

伦理声明

本研究涉及从叙述文本中自动生成问答对，并结合难度级别和叙述元素等控制属性。使用的数据集 FairytaleQA 由公开可用的童话故事

中由人类创作的问答对组成。不包含任何可识别个人身份或敏感信息，确保遵守数据使用的伦理指南。对生成的问答对进行了自动化指标评估和人工检查，以识别潜在错误，如幻觉内容和无意义问题。我们承认这些模型可能引入非预期的错误或偏见。尽管本文不侧重于错误缓解，但建议未来的工作可以探索扩展的人为参与验证，以增强生成问答对的可靠性，特别是在部署场景中。

10

致谢

作者感谢 Masaki Uto 教授和 Yuto Tomikawa 在相关前期工作中提供的有益澄清和讨论。该工作由 UID/00027——人工智能与计算机科学实验室（LIACC）提供资金支持，该实验室得到由 Funda ç ão para a Ciência e a Tecnologia, I.P./ MCTES 通过国家资金资助。Bernardo Leite 获得由 FCT 资助的博士奖学金（参考号 2021.05432.BD）。

References

A 叙事控制：语义相似性

表 6 展示了从叙事控制角度，通过 BLEURT (**?**) 测量的结果。其目的是显示在引入叙事控制属性时，与真实问题的语义相似性得到改善。正如用 ROUGE $_L$ -F1 相似性观察到的（回忆第 5.2 节），此趋势在所有七个叙事标签中均有体现。当叙事和难度标签融合时，我们观察到类似的改善趋势，与仅引入叙事属性相当。这些结果进一步支持假设 1 (H1) ——引入叙事属性将生成的问题与真实问题更相似——这表明我们的方法可以控制生成问题中的叙事元素。

B 仅难度与难度加叙事控制

为了比较仅基于难度操作与结合难度和叙述属性时的难度控制情况，图 6 提供了两种设置在各个级别的性能概述。两种设置均显示出预期的趋势：正确答案的百分比随着难度的增加而减少。然而，对观察到的数据点进行线性近似可以揭示，当两个属性结合时，这种减少不那么显著，但整体仍然一致。

C 受控影响的其他语言特征

表格 7 展示了生成的问答对中的平均单词数量。在生成的答案中，只有难度标签被纳入时，没有观察到显著的趋势。在生成的问题中，虽然观察到一个上升趋势，但并不显著。当叙事和难度标签结合时，没有观察到任何趋势。基

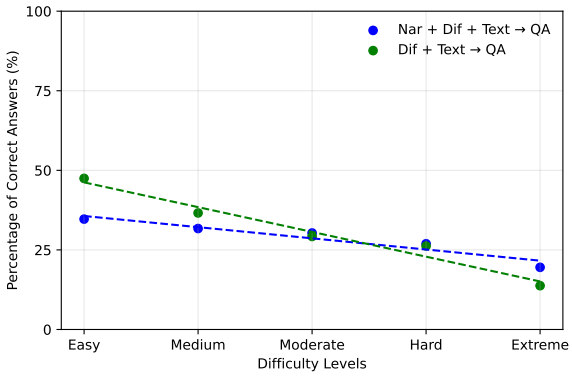

Figure 6: 各难度等级的正确答案百分比。

于这些发现，我们得出结论，在我们的实验中，生成的问答对的平均长度不受难度或叙事控制标签影响。图 7 说明了生成问题中初始疑问词的比例。仅使用难度标签时（上图），较高难度水平显示出像“为什么”和“怎么”这样的词的增加，以及像“什么”、“谁”和“哪里”这样的词的减少。这与预期一致，因为“为什么”和“怎么”通常与需要更高认知努力的问题相关，正如布鲁姆的分类学所述 (**?**) 。当叙事和难度标签融合时（下图），所有疑问词的比例在难度水平上更为一致。这一结果是预期的，因为这种设置旨在控制难度水平，同时也要求某些叙事元素。在这种情况下，叙事标签主要影响疑问词的选择（例如，“谁”用于与角色相关的问题），而不是难度标签。

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: 叙事控制：在测试集中，生成问题与真实问题在叙事元素上的语义相似度（BLEURT）。文本 → QA 被用作基准，以评估叙事控制是否有助于生成问题接近真实问题。

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: 生成问题（Q）和答案（A）的平均字数。

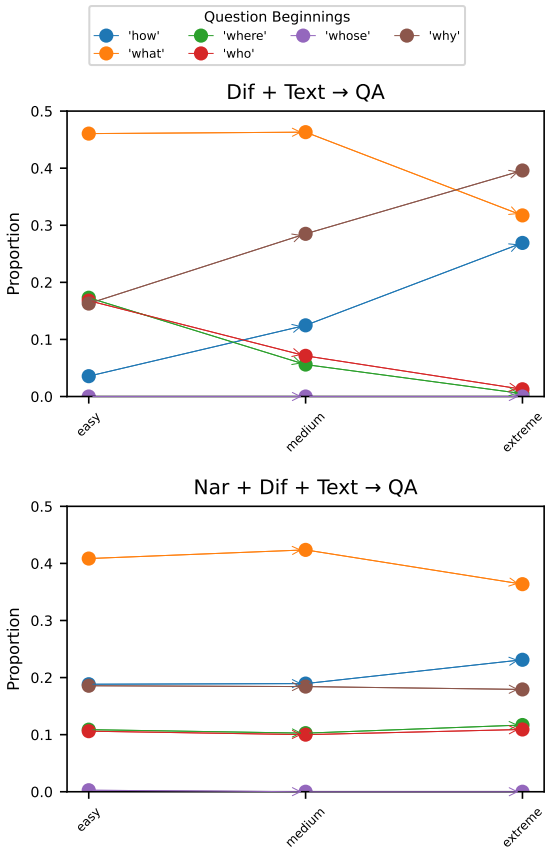

Figure 7: 生成问题中初始疑问项的比例（箭头线表示增加/减少趋势）。